DP-900 · CHEAT SHEET · CERTIFYDECK

# Microsoft Certified: Azure Data Fundamentals

Dominios, pesos oficiales y habilidades clave del temario de DP-900, para repasar antes del examen.

Nivel: Beginner  Duración: 45 min  Nota de corte: 700/1000  Fuente: guía de estudio oficial de Microsoft Learn

## 1. Describe core data concepts

25–30%

### Describe ways to represent data

- Los datos estructurados siguen un esquema fijo, de modo que todas las instancias de una entidad tienen los mismos campos. Lo más habitual es representarlos en tablas: las filas son las instancias y las columnas, los atributos, y cada columna guarda un tipo de datos concreto. Suelen almacenarse en bases de datos relacionales, donde las tablas se enlazan mediante claves.
- Los datos semiestructurados tienen cierta estructura pero admiten variaciones entre instancias: por ejemplo, un cliente puede tener varios correos y otro ninguno. JSON es un formato muy común, con un esquema jerárquico donde un atributo puede ser un objeto o una colección. JSON es solo una de las formas posibles, y las bases de datos de documentos guardan documentos JSON.
- Los datos no estructurados no tienen una estructura concreta: documentos, imágenes, audio, vídeo y archivos binarios. Se almacenan como binario sin procesar que las aplicaciones deben interpretar, y los profesionales de datos los llaman BLOB. Azure Blob Storage está optimizado para guardar enormes cantidades de este tipo de datos. Los datos vectoriales (incrustaciones) permiten a la IA responder sobre ellos.

### Identify options for data storage

- El formato de archivo depende del tipo de datos, de las aplicaciones que los usan y de si deben ser legibles o estar optimizados. El texto delimitado (CSV, TSV, ancho fijo) sirve para datos estructurados; JSON, para estructurados y semiestructurados; XML es legible pero más verboso. Parquet (columnar) y Avro (por filas) comprimen y procesan mejor, y Delta Lake añade transacciones sobre Parquet.
- Una base de datos es un sistema dedicado a gestionar registros de datos, no solo archivos. Las relacionales guardan datos estructurados en tablas con claves principales, normalizadas y consultadas con SQL. Las no relacionales, a menudo llamadas NoSQL, no aplican un esquema relacional: clave-valor, documentos, familia de columnas y grafos. Ningún almacén sirve para todo, por eso se combinan modelos.
- Cada necesidad tiene su almacén: Azure Blob Storage guarda datos no estructurados (imágenes, vídeo, copias de seguridad) y, con espacio de nombres jerárquico, forma el lago de datos; Azure Files ofrece recursos compartidos SMB/NFS; Azure Table storage guarda datos clave/atributo sin esquema; Azure Cosmos DB es la base NoSQL de baja latencia y alcance global; y la familia Azure SQL cubre los datos relacionales.

### Describe common data workloads

- Una carga transaccional (OLTP) registra eventos de negocio como unidades de trabajo pequeñas y rápidas, sobre todo escrituras con lecturas moderadas, y da soporte a aplicaciones de línea de negocio. Aplica las operaciones CRUD con semántica ACID: atomicidad, coherencia, aislamiento y durabilidad. Sus datos suelen estar muy normalizados y con esquema impuesto.
- Una carga analítica usa sistemas de solo lectura o de lectura mayoritaria con grandes volúmenes de datos históricos. Los datos se cargan con ETL o ELT en un lago de datos, un lakehouse o un almacén de datos, y pueden agregarse en un modelo OLAP o semántico para informes y paneles. Microsoft Fabric y Azure Databricks son plataformas de análisis modernas.

### Identify roles and responsibilities for data workloads

- El administrador de bases de datos diseña, implementa y mantiene sistemas de bases de datos locales y en la nube. Garantiza su disponibilidad, rendimiento y optimización, define los planes de copia de seguridad y recuperación y gestiona la seguridad, concediendo o denegando el acceso. En servicios PaaS como Azure SQL Database deja de gestionar el motor, el sistema operativo y el hardware.
- El ingeniero de datos diseña e implementa canalizaciones de ingesta, actividades de limpieza y transformación y almacenes de datos para cargas analíticas. Gestiona y supervisa las canalizaciones, mantiene la privacidad de los datos entre local y nube y trabaja con herramientas como Azure Data Factory, Microsoft Fabric, Azure Databricks, Azure Stream Analytics y Microsoft Purview.
- El analista de datos explora los datos para identificar tendencias y relaciones, diseña modelos analíticos y convierte los datos sin procesar en información útil mediante informes y visualizaciones, según los requisitos del negocio. Suele usar Power BI. Se diferencia del administrador de bases de datos (permisos, copias de seguridad) y del ingeniero de datos (canalizaciones y limpieza), y una misma persona puede ejercer varios roles.

## 2. Identify considerations for relational data on Azure

20–25%

### Describe relational concepts

- En una base de datos relacional los datos son tabulares: las entidades son tablas, cada fila es una instancia y cada columna una propiedad, y todas las filas de una tabla tienen las mismas columnas. La clave principal identifica de forma única cada fila y la clave externa hace referencia a la clave principal de otra tabla, lo que mantiene las relaciones y la integridad referencial. Encaja bien con cargas OLTP.
- La normalización organiza los datos en tablas relacionadas según reglas (formas normales) que eliminan la redundancia y las dependencias incoherentes. Así un dato, como la dirección de un cliente, se guarda en un solo lugar y se cambia una vez. La tercera forma normal suele bastar; el coste es que se necesitan más tablas, y demasiadas tablas pequeñas pueden afectar al rendimiento.
- SQL permite crear tablas (CREATE TABLE), insertar (INSERT), modificar (UPDATE) y eliminar filas (DELETE) y consultar datos (SELECT), con WHERE para filtrar y JOIN para combinar tablas. Las instrucciones DDL (CREATE, ALTER, DROP) definen estructuras de datos y las DML (INSERT, UPDATE, DELETE, SELECT) trabajan con las filas. Cada instrucción es una unidad atómica: se completa o falla por entero.
- Una vista es una tabla virtual definida por una consulta. Un índice es una copia ordenada de una columna que acelera las consultas pero hay que mantenerlo en cada inserción, actualización o eliminación. Un procedimiento almacenado agrupa instrucciones Transact-SQL con parámetros y un desencadenador (trigger) se ejecuta solo ante INSERT, UPDATE o DELETE. Las claves principales y externas son restricciones que garantizan la integridad.

### Describe relational Azure data services

- Azure SQL es una familia de productos basados en el motor de SQL Server. Azure SQL Database es PaaS para aplicaciones nuevas en la nube, con bases de datos únicas o grupos elásticos. Azure SQL Managed Instance es PaaS con compatibilidad casi total con SQL Server, ideal para migrar instancias con cambios mínimos. SQL Server en Azure Virtual Machines es IaaS: da control total del sistema operativo y del motor, pero el cliente los administra.
- Azure Database for MySQL y Azure Database for PostgreSQL son servicios PaaS basados en las ediciones comunitarias de esos motores, así que las aplicaciones existentes migran con pocos cambios. Azure aplica parches, crea copias de seguridad automáticas y ofrece alta disponibilidad, aunque algunas funciones de servidor, como ciertas extensiones, no están disponibles. Se puede detener el servidor para ahorrar costes.

## 3. Describe considerations for working with non-relational data on Azure

15–20%

### Describe the capabilities of Azure storage

- Azure Blob Storage es el almacenamiento de objetos de Azure para grandes volúmenes de datos no estructurados: imágenes, vídeo, copias de seguridad y análisis. Los blobs viven en contenedores dentro de una cuenta de almacenamiento y pueden ser en bloques, en anexos o en páginas. Los niveles de acceso (frecuente, esporádico, frío y archivo) equilibran coste y latencia, y la administración del ciclo de vida automatiza los cambios de nivel.
- Azure Files ofrece recursos compartidos de archivos totalmente administrados accesibles por SMB y NFS, y también por API REST. Se pueden montar a la vez desde clientes en la nube y locales, lo que permite sustituir servidores de archivos o trasladar aplicaciones sin cambiar el código. Azure File Sync usa Windows Server como caché local de un recurso compartido de Azure, y la redundancia se define a nivel de cuenta de almacenamiento.
- Azure Table storage es un almacén NoSQL de clave/atributo con diseño sin esquema para datos estructurados no relacionales, como datos de usuario o información de dispositivos. Una tabla es una colección de entidades, y cada entidad es un conjunto de propiedades con una clave de partición y una clave de fila. No admite combinaciones complejas ni claves externas; Azure Cosmos DB for Table añade distribución global e índices secundarios.

### Describe the capabilities and features of Azure Cosmos DB

- Azure Cosmos DB es una base de datos NoSQL totalmente administrada, con esquema flexible, indexación automática, escalado elástico, baja latencia y distribución global. Encaja en IoT, catálogos de productos, juegos, aplicaciones web y móviles y personalización, y su change feed permite arquitecturas por eventos. No sirve para analítica (OLAP), donde se usa Microsoft Fabric, ni para apps muy relacionales, donde se usa Azure SQL.
- Al crear una cuenta de Azure Cosmos DB se elige la API, que determina el formato de datos, el lenguaje de consulta y las bibliotecas cliente. NoSQL es la API nativa (documentos JSON, sintaxis tipo SQL) y la recomendada para aplicaciones nuevas; MongoDB (BSON, MQL), Cassandra (familias de columnas, CQL) y Table (clave-valor) facilitan migrar aplicaciones existentes, y Gremlin sirve para grafos de vértices y aristas.

## 4. Describe an analytics workload

25–30%

### Describe common elements of large-scale analytics

- La ingesta y el procesamiento cargan datos de bases transaccionales, archivos y flujos en tiempo real en un lago de datos o un almacén de datos, por lotes o en streaming. En ETL los datos se transforman antes de cargarse; en ELT se copian y se transforman en el destino. En Microsoft Fabric se ingiere con Data Factory, accesos directos, mirroring, Eventstream o cuadernos, y en Azure Databricks con Lakeflow o cuadernos.
- Un almacén de datos es una base relacional con esquema optimizado para analítica (tablas de hechos y dimensiones, esquema en estrella o de copo de nieve). Un lago de datos es un almacén de archivos con esquema en lectura que admite datos estructurados, semiestructurados y no estructurados. El lakehouse combina ambos con Delta Lake. En Azure: Fabric Lakehouse, Fabric Warehouse y Azure Databricks.
- Microsoft Fabric es una plataforma de analítica SaaS unificada cuyas cargas de trabajo (Data Factory, Data Engineering, Data Warehouse, Data Science, Real-Time Intelligence y Power BI) comparten OneLake, un lago de datos único por inquilino. Azure Databricks es una plataforma basada en Apache Spark que usa Delta Lake como formato nativo e incluye cuadernos, Unity Catalog para la gobernanza y Databricks SQL.

### Describe considerations for real-time data analytics

- El procesamiento por lotes recoge y almacena registros y los procesa juntos en una sola operación, con latencias de horas y adecuado para análisis complejos. El procesamiento de flujos trata cada evento al llegar, con latencias de segundos o milisegundos, y sirve para agregados sencillos y operaciones críticas en el tiempo. Muchas soluciones los combinan, por ejemplo con la arquitectura lambda.
- Microsoft ofrece varias opciones de tiempo real. Microsoft Fabric Real-Time Intelligence incluye Eventstreams, Eventhouse (consultas KQL), paneles en tiempo real y Activator; Azure Stream Analytics es un servicio PaaS que ejecuta consultas perpetuas; y Spark Structured Streaming funciona en Fabric y Azure Databricks. Los eventos se ingieren con Azure Event Hubs o Azure IoT Hub.

### Describe data visualization in Microsoft Power BI

- Power BI es un conjunto de herramientas y un workload de Microsoft Fabric. En Power BI Desktop, una aplicación de Windows, se importan datos, se modelan y se crean informes interactivos; el servicio Power BI publica, comparte y programa actualizaciones; la aplicación para teléfonos consume contenido. Report Builder crea informes paginados e imprimibles.
- Un modelo semántico de Power BI se construye con tablas relacionadas: las tablas de hechos guardan las medidas numéricas y las de dimensiones las entidades por las que se agrupa, en un esquema de estrella o de copo de nieve. Las relaciones propagan filtros, las jerarquías permiten profundizar y los datos se almacenan en memoria con el motor VertiPaq.
- Cada visualización responde a una necesidad: tablas y tarjetas para valores y cifras clave, columnas o barras para comparar categorías, líneas para tendencias, circulares para proporciones de un total, dispersión para correlaciones entre dos medidas y mapas para ubicaciones. En Power BI los elementos visuales de un informe están vinculados y se filtran entre sí.