Microsoft Azure Data Fundamentals

Conceptos básicos de datos

- Conceptos principales de los datos
- Roles y servicios de datos

Data relacional en Azure

- Conceptos fundamentales de datos relacionales
- Servicios de base de datos relacionales en Azure

Data no-relacional en Azure

- Azure Storage para datos no relacionales
- Conceptos de Azure Cosmos DB

Análisis de datos en Azure

- Fundamentos del almacenamiento de datos moderno
- Fundamentos de la analítica en tiempo real
- Fundamentos de la visualización de datos

Conceptos básicos de datos

Conceptos principales de los datos

- Los datos son una colección de elementos (números, descripciones y observaciones).
- Registran información.
- Las estructuras de datos en las que se organizan suelen representar **entidades** de una organización (clientes,productos,pedidos de ventas, etc).
- Los datos se clasifican en **estructurados**, **semiestructurados** o **no estructurados**.

Datos estructurados

- Se ajustan a un **esquema** fijo (comparten mismos campos).
- Esquema **tabular** (en tablas).
- Cada **fila** representa una **instancia de una entidad**.
- Se suelen almacenar en una **base de datos** en un modelo **relacional**

CustomerID	Title	FirstName	MiddleName	LastName	Suffix	CompanyName	Phone
1	Mr.	Orlando	N.	Gee	NULL	A Bike Store	245-555-0173
2	Mr.	Keith	NULL	Harris	NULL	Progressive Sports	170-555-0127
3	Ms.	Donna	F.	Carreras	NULL	Advanced Bike Components	279-555-0130
4	Ms.	Janet	M.	Gates	NULL	Modular Cycle Systems	710-555-0173
5	Mr.	Lucy	NULL	Harrington	NULL	Metropolitan Sports Supply	828-555-0186
6	Ms.	Rosmarie	J.	Carroll	NULL	Aerobic Exercise Company	244-555-0112
7	Mr.	Dominic	P.	Gash	NULL	Associated Bikes	192-555-0173
10	Ms.	Kathleen	M.	Garza	NULL	Rural Cycle Emporium	150-555-0127
11	Ms.	Katherine	NULL	Harding	NULL	Sharp Bikes	926-555-0159
12	Mr.	Johnny	A.	Caprio	Jr.	Bikes and Motorbikes	112-555-0191
16	Mr.	Christopher	R.	Beck	Jr.	Bulk Discount Store	1 (11) 500 555-0132
18	Mr.	David	J.	Liu	NULL	Catalog Store	440-555-0132
19	Mr.	John	A.	Beaver	NULL	Center Cycle Shop	521-555-0195
20	Ms.	Jean	P.	Handley	NULL	Central Discount Store	582-555-0113
21	Mr.	Jinghao	NULL	Liu	NULL	Chic Department Stores	928-555-0116
22	Ms.	Linda	E.	Burnett	NULL	Travel Systems	121-555-0121
23	Mr.	Kerim	NULL	Hanif	NULL	Bike World	216-555-0122
24	Mr.	Kevin	NULL	Liu	NULL	Eastside Department Store	926-555-0164
25	Mr.	Donald	L.	Blanton	NULL	Coalition Bike Company	357-555-0161
28	Ms.	Jackie	E.	Blackwell	NULL	Commuter Bicycle Store	972-555-0163
29	Mr.	Bryan	NULL	Hamilton	NULL	Cross-Country Riding Supp...	344-555-0144
30	Mr.	Todd	R.	Logan	NULL	Cycle Merchants	783-555-0110
34	Ms.	Barbara	J.	German	NULL	Cycles Wholesaler & Mfg.	1 (11) 500 555-0181
37	Mr.	Jim	NULL	Geist	NULL	Two Bike Shops	724-555-0161

```
// Customer 1
{
  "firstName": "Joe",
  "lastName": "Jones",
  "address":
  {
    "streetAddress": "1 Main St.",
    "city": "New York",
    "state": "NY",
    "postalCode": "10099"
  },
  "contact":
  [
    {
      "type": "home",
      "number": "555 123-1234"
    },
    {
      "type": "email",
      "address": "joe@litware.com"
    }
  ]
}
```

Datos semiestructurados

- Tienen cierta estructura, pero permiten **variaciones** entre las **instancias de entidad**.
- Formato común es la **notación de objetos JavaScript** (JSON)

Datos no estructurados

- Documentos, imágenes, datos de audio y de vídeo y archivos binarios

Almacenes de datos

- Los datos se almacenan sin importar su clasificación
- Se espera que puedan ser recuperados para su análisis y generación de informes
- Dos categorías generales:
  - Almacenes de archivos
  - Bases de datos

Los datos son la base sobre la que se crea todo el sofotware. Implica conocer formatos de datos comunes, cargas de trabajo, roles y servicios

Conceptos básicos de datos

Almacenamiento de archivos

- Capacidad de almacenar es un elemento básico de un sistema informático.
- Los archivos de datos importantes se almacenan en un **sistema de almacenamiento de archivos compartido**.
- El formato de archivo que se usa para almacenar datos depende de varios factores:
  - Tipo de datos que se almacenan (estructurados,semi...).
  - Las aplicaciones y servicios que leeran, escribiran y procesarán los datos.
  - Los archivos de datos deberan ser legibles u optimizados para el almacenamiento.

Formato de archivos comunes

Archivos de texto delimitado

- Campos separados por comas (generalmente) y las filas finalizan con una nueva linea (\n)
- Primera linea pueden ser las columnas

Notación de objetos JavaScript (JSON)

- Esquema de documento jerárquico para definir objetos que tienen varios atributos
- Objetos se incluyen entre llaves ({..}) y las colecciones de estos entre corchetes ([..])
- Los atributos se representan mediante pares **nombre:valor** y se separan por comas

Lenguaje de marcado extensible (XML)

- Mismo objetivo que JSON
- Usa etiquetas para representar elementos y atributos

Objeto binario grande (BLOB)

- En última instancia se almacenan como datos binarios
- No tienen formato legible pues no se asignaron según un esquema de codificación de caracteres

Formato de archivos optimizados

- Optimizados para el procesamiento y el espacio de almacenamiento

Avro

- Formato basado en filas
- Cada registro contiene un encabezado de la estructura de los datos del mismo en JSON y los datos en binario
- Adecuado para comprimir datos y reducir requisitos de almacenamiento y de ancho de banda de red

ORC

- Formato de columnas de filas optimizadas (organiza los datos en columnas)
- Optimiza operaciones de lectura y escritura
- Contiene franjas de datos. Cada una contiene los datos de un conj. de columnas
- Una franja contiene un índice de las filas de dicha franja, los datos y un pie de página con información estadística

Parquet

- Formato basado en columnas y cada archivo parquet contiene grupos de filas
- Los datos de cada columna se almacenan juntos en el mismo grupo de filas
- Destaca en almacenar y procesar tipos de datos anidados

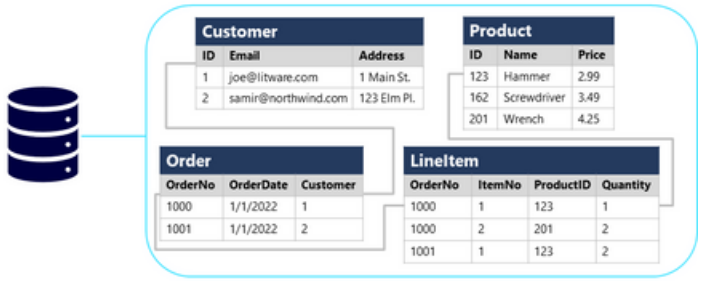
Conceptos básicos de datos

Bases de datos

- Definen un sistema central en el que los datos se almacenan y se pueden consultar.

Bases de datos relacionales

- Almacenan y se pueden consultar datos estructurados
- Los datos se almacenan en tablas que representan entidades
- Cada instancia de una entidad se le asigna una clave principal única
- El uso de claes permite normalizar una base de datos relacional
- Eliminación de valores de datos duplicados
- Se administran y consultan mediante SQL



Customer

ID	Email	Address
1	joe@litware.com	1 Main St.
2	samir@northwind.com	123 Elm Pl.

Product

ID	Name	Price
123	Hammer	2.99
162	Screwdriver	3.49
201	Wrench	4.25

Order

OrderNo	OrderDate	Customer
1000	1/1/2022	1
1001	1/1/2022	2

LineItem

OrderNo	ItemNo	ProductID	Quantity
1000	1	123	1
1000	2	201	2
1001	1	123	2

Bases de datos no relacionales

- No aplican un esquema relacional a los datos
. Conocidas como bases de datos NoSQL, pero algunas admiten variantes del lenguaje SQL
- 4 tipos comunes

Clave-valor

- Cada registro consta de una clave única y un valor asociado que está en cualquier formato

Products

Key	Value
123	"Hammer ($2.99)"
162	"Screwdriver ($3.49)"
201	"Wrench ($4.25)"

De documentos

- Forma específica de clave-valor
- El valor es un documento JSON
- El sistema está optimizado para analizar y consultar

Customers

Key	Document
1	{ "name": "Joe Jones", "email": "joe@litware.com" }
2	{ "name": "Samir Nadoy", "email": "Samir@northwind.com" }

Familia de columnas

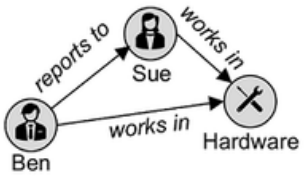
- Datos tabulares con filas y columas
- Columnas divididas en grupos

Orders

Key	Customer		Product	
	Name	Address	Name	Price
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

Grafos

- Almacenan entidades como nodos con vínculos para definir relaciones entre ellas



Conceptos básicos de datos

Procesamiento de datos transaccionales

- Un sistema transaccional registra las **transacciones** que encapsulan eventos de la organización.
- Un sistema de procesamiento de datos transaccional es la función principal de la informática empresarial.
- Una transacción es una **unidad de trabajo** pequeña y discreta (Ejemplo: movimiento de dinero)
- Estos sistemas son de gran volumen y se debe acceder a sus datos con rapidez
- El trabajo que realizan estos sistemas se conoce como **procesamiento de transacciones en línea** (OLTP)

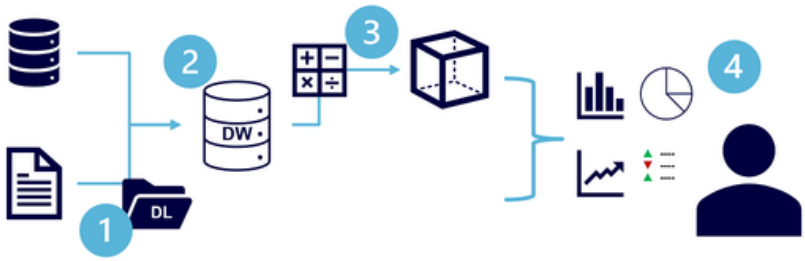
Soluciones OLTP

- Sistema de base de datos en el que el almacenamiento está optimizado para lectura y escritura de datos (cargas de trabajo con operaciones CRUD).
- Admiten transacciones que admiten la semántica ACID:
  - **Atomicidad** : Cada transacción es única. Deben completarse correctamente todos sus pasos o produce un error general.
  - **Coherencia** : Las transacciones solo pueden admitir estados validos en la base de datos.
  - **Aislamiento** : Las transacciones simultáneas no pueden interferir entre sí y deben devolver estados coherentes en la base de datos
  - **Durabilidad** : Una transacción confirmada debe permanecer como tal (Persistencia ante fallos).
- Sistema OLTP se usan para admitir aplicaciones activas que procesan datos empresariales denominadas de **línea de negocio** (LOB)

Procesamiento de datos analíticos

- Sistema de **principalemente solo lectura**.
- Generalmente almacena grandes volúmenes de datos históricos o métricas empresariales.
- Los análisis se basan en una instantánea de los datos o una serie de estas.

Arquitectura común para el análisis a escala empresarial



1. Los archivos de datos se almacenan en un **lago de datos** central
2. Un **proceso ETL** permite copiar datos de **archivos y bases de datos OLTP** en un **almacenamiento optimizado en lectura** con un esquema basado en **tablas de hechos y dimensiones**

3. Los datos del almacenamiento de datos se pueden cargar en un **modelo de procesamiento analítico en línea (OLAP) o un cubo**
4. Los datos del **lago de datos, almacenamiento de datos y el modelo analítico** se pueden **consultar** para generar informes, visualizaciones y paneles

Lagos de datos

- Comunes en estos escenarios
- Recopilan y analizan gran volumen de datos basados en archivos

Almacenamiento de datos

- Almacenan datos en esquema relacional y optimizado para lectura
- Suelen requerir alguna desnormalización de los datos

Modelo OLAP

- Tipo agregado de almacenamiento de datos
- Optimizado para cargas de trabajo analíticas debido a sus modelos

Conceptos básicos de datos

Roles de trabajo en el mundo de los datos

- Amplia variedad de roles implicados en administración, control y uso de datos
- Orientados a negocios, a ingeniería, investigación o hibridos

Roles de trabajo principales que se ocupan de los datos de la mayoría de las organizaciones

Administrador de base datos

- Responsable del diseño, implementación, mantenimiento y aspectos operativos de los sistemas de bases de datos
- Responsables de la disponibilidad general, optimizaciones y rendimiento coherente de las bases de datos
- Realizan copias de seguridad trabajando con las partes interesadas
- Realizan planes de recuperación que permiten reponerse tras un desastre natural
- Administran la seguridad de los datos en la base de datos
- Conceder privilegios sobre los datos
- Conceden o deniegan el acceso a los usuarios según corresponda

Ingenierio de datos

- Colaboran con las partes interesadas para diseñar e implementar cargas de trabajo relacionado con datos
- Canalizaciones de ingesta de datos, actividades de limpieza y transformación
- Uso de almacenes de datos para cargas de trabajo analíticas
- Uso de amplia gama de tecnología de plataforma de datos, bases de datos relacionales y no relacionales, almacenes de archivos y flujos de datos
- Garantizan la privacidad de datos durante su estadia en reposo y cuando se transporta
- Administra y supervisa las canalizaciones de datos para asegurarse que las cargas de datos funcionen

Analista de datos

- Maximiza el valor de sus recursos de datos
- Exploran datos para identificar tendencias y relaciones, diseñar e implementar modelos analíticos
- Habilitan funcionalidades de análisis avanzado mediante informes y visualizaciones
- Ocupan el procesamiento de datos para convertirlos en información pertinente

Tipos de análisis aplicado a los datos

- Descriptivo : ¿Qué está sucediendo?
- Diagnostico : ¿Por qué está pasando?
- Predictivo : ¿Que sucederá?
- Prescriptivo : ¿Qué acciones debemos tomar para lograr esto?

Existe una última que es "Cognitivo", aplica los servicios cognitivos a los datos

Conceptos básicos de datos

Servicios de datos en Azure

Se explorarán algunos de los servicios de datos más usados para soluciones transaccionales y analíticas modernas.

Azure SQL

- Nombre colectivo de una familia de soluciones de base de datos relacionales basadas en el motor de SQL Server
- Servicios específicos incluyen:
  - **Azure SQL Database :** Base de datos como PaaS .
  - **Azure SQL Managed Instance :** Instancia hospedada de SQL Server con más configuración y responsabilidad administrativa.
  - **Azure SQL Virtual Machine :** Máquina virtual con una instalación de SQL Server. Capacidad de configuración máxima y responsabilitad total del servidor y sistema operativo.

NOTA:
- Los administradores de BD aprovisionan bases de datos para **admitir aplicaciónes LOB** que necesitan almacenar datos transaccionales.
- Los ingenierios de datos **usan BD como orígenes para canalizaciones de datos que realizan operaciones ETL** para ingerir los datos transaccionales en un sistema analítico.

Azure Database para bases de datos relacionales de código abierto

- Servicios administrados para sistemas populares de bases de datos:
- **Azure Database for MySQL :** Sistema de administración de BD. Util para la pila LAMP.
- **Azure Database for MariaDB :** Motor de base de datos reescrito y optimizado respecto a MySQL.
- **Azure Database for PostgreSQL :** Base de datos híbrida de objetos relacionales.

Azure Cosmos DB

- Sistema de base de datos no relacional (NoSQL) a escala global.
- Admite varias interfaces de programación de aplicaciones (API).
- Permite almacenar y administrar datos como documentos JSON, pares clave-valor, familia de columnas y grafos.

NOTA:
- Los administradores de BD pueden aprovisionar estas instancias, pero los **desarrolladores de software son quienes administrar el almacenamiento de datos NoSQL** como parte de la arquitectura general de la aplicación.

Azure Storage

Permite almacenar datos en:
- **Contenedores de blobs :** Almacenamiento escalable para archivos binarios.
- **Recursos compartidos de archivos :** Compartidos en red (Ejemplos : discos compartidos en red).
- **Tablas :** Almacenamiento de clave-valor para lectura y escritura veloz.

NOTA:
- Los ingenieros de datos usan Azure Storage para hospedar lagos de datos, con un espacio de nombres jerárquico.

Azure Data Factory

- Define y programa canalizaciones de datos para transferir y transformar datos.
- Puede integrar las canalizaciones con otros servicios de Azure.

NOTA:
- Los ingenierios de datos usan Data Factory para **compilar soluciones de ETL** que rellenen almacenes de datos analíticos con datos de sistemas transaccionales de toda la organización.

Conceptos básicos de datos

Servicios de datos en Azure

Azure Synapse Analytics

Solución completa y unificada de análisis de datos que proporciona una interfaz de servicio única para varias funcionalidades:

- **Pipelines :** Basada en la misma tecnología que **Azure Data Factory**
- **SQL :** Motor de base de datos SQL altamente escalable, optimizado para cargas de trabajo de almacenamiento de datos
- **Apache Spark :** Sistema de procesamiento de datos distribuidos
- **Azure Synapse Data Explorer :** Solución de análisis de datos de alto rendimiento que está optimizada para consultas en tiempo real de datos de registro y telemetría mediante el Lenguaje de consulta Kusto (KQL).

NOTA:
- Los ingenierios de datos usan Synapse Analytics para crear una solución de análisis de datos unificada que **combina canalizaciones de ingesta, almacenamiento en almacén de datos y en lagos de datos en un solo servicio**.
- Los analistas de datos pueden usar grupos de **Spark y SQL** y aprovechar la integración con servicios como **Azure Machine Learning y Microsoft Power BI**

Azure Databricks

Versión integrada de Azure de la plataforma Databricks

NOTA:
- Los ingenierios de datos usan para crear almacenes de datos analíticos en Azure Databricks

Azure HDInsight

Proporciona clústeres hospedados en Azure para tecnologías de procesamiento de macrodatos de código abierto de Apache:

- **Apache Spark :** Sistema de procesamiento de datos distribuidos
- **Apache Hadoop :** Sistema distribuidos que usa trabajos de MapReduce
- **Apache HBase :** Sistema para consultas y almacenamiento de datos NoSQL a gran escala
- **Apache Kafka :** Agente de mensajes para el procesamiento de flujos de datos
- **Apache Storm :** Procesamiento de datos en tiempo real mediante topología de spouts y bolts

Azure Stream Analytics

Motor de procesamiento de flujos en tiempo real que captura un flujo de datos de una entrada, aplica una consulta para extraer y manipular los datos y escribe los resultados en una salida para su análisis o procesamiento posterior.

- Se usa para capturar datos de streaming para su ingesta o visualización en tiempo real

Azure Data Explorer

Servicio independiente que permite consultar datos de telemetría y del registro

- Se usa para consultar y analizar datos con atributos de marca de tiempo habituales en archivos de registro y datos de telemetría de IoT

Azure Purview

Solución para la gobernanza y la detectabilidad de datos de toda la empresa, para crear mapa de los datos y realizar un seguimiento del linaje de datos

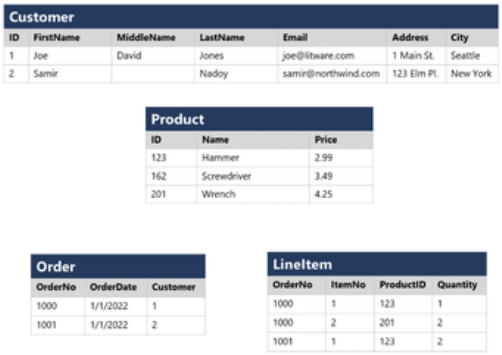
Microsoft Power BI

Plataforma para el modelado de datos analíticos y elaboración de informes

Datos relacionales en Azure

Fundamentos de datos relacionales

Datos relacionales



Customer

ID	FirstName	MiddleName	LastName	Email	Address	City
1	Joe	David	Jones	joe@litware.com	1 Main St.	Seattle
2	Samir		Nadoy	samir@northwind.com	123 Elm Pl.	New York

Product

ID	Name	Price
123	Hammer	2.99
162	Screwdriver	3.49
201	Wrench	4.25

Order

OrderNo	OrderDate	Customer
1000	1/1/2022	1
1001	1/1/2022	2

LineItem

OrderNo	ItemNo	ProductID	Quantity
1000	1	123	1
1000	2	201	2
1001	1	123	2

- Se encuentran en una base de datos relacional y se modelan en forma de tablas
- Una tabla (relacional) es un formato para datos estructurados y cada fila de una tabla tiene las mismas columnas
- No todas las columnas necesitan tener un valor (NULL)
- Los tipos de datos disponibles que se pueden usar al definir las columnas de una tabla dependen del sistema de base de datos, aunque hay tipos de datos estándar definidos por ANSI

Comprensión de la normalización

- Proceso de diseño de esquemas que reduce al mínimo la duplicación de los datos e impone la integridad de estos.
- Hay muchas reglas complejas que definen el proceso de refactorización de los datos en varios niveles (o formas) de normalización

Exploración de SQL

- Lenguaje estándar para los sistemas de administración de bases de datos relacionales.
- Tipos de instrucción SQL agrupados en grupos lógicos principales
  - **Lenguaje de definición de datos (DDL)**
  - **Lenguaje de control de datos (DCL)**
  - **Lenguaje de manipulación de datos (DML)**

Instrucciones DDL

- Crear, modificar y quitar tablas y otros objetos de una base de datos (tabla, procedimientos almacenados, vistas, etc.)

Instrucciones DCL

- Administrar el acceso a objetos de una base de datos mediante la concesión, denegación o revocación de permisos a usuarios o grupos específicos

Instrucciones DML

- Insertar, modificar o eliminar las filas de las tablas

Descripción de objetos de base de datos

Vista

Tabla virtual basada en los resultados de una consulta SELECT.

Procedimientos almacenados

- Define instrucciones SQL que se pueden ejecutar a petición
- Encapsulan lógica de programación en una base de datos

Índices

- Ayuda a buscar datos en una tabla
- Crea una estructura basada en árbol que el optimizador de consultas del sistema de BD usa para buscar
- Son costosos de crear, por lo que se recomienda su uso en casos puntuales

Datos relacionales en Azure

Servicios y capacidades de Azure SQL

SQL Server Virtual Machine	Azure SQL Managed Instance	Azure SQL Database
Tipos de servicio en la nube		
IaaS	**PaaS**	**PaaS**
Compatibilidad con SQL Server		
- Totalmente compatible con instalaciones físicas y virtualizadas locales - Migración fácil usando lift-and-shift sin cambios	- Casi completamente compatible con SQL Server - Migración con cambios mínimos en el código mediante el servicio Azure Database Migration	- Admite la mayoria de funcionalidades básicas de SQL Server - Es posible que algunas características no estén disponibles para la migración
Arquitectura		
- Las instancias de SQL Server se instalan en una máquina virtual - Cada instancia puede admitir varias bases de datos	- Cada instancia administrada puede admitir varias bases de datos - Grupos de instancias para compartir recursos de forma eficaz entre instancias más pequeñas	- Aprovisiona una base de datos única en un servidor dedicado y administrado (lógico) - Grupo elástico para compartir recursos entre bases de datos
Disponibilidad		
- Disponibilidad de 99,99%	- Disponibilidad de 99,99%	- Disponibilidad de 99,995%
Administración		
- Administra todos los aspectos del servidor y sistema operativo: configuración, copias de seguridad y otras tareas de mantenimiento	- Las actualizaciones, copias de seguridad y recuperación están automatizadas	- Las actualizaciones, copias de seguridad y recuperación están automatizadas
Adicionales		
- Escalar verticalmente la plataforma en la que se ejecuta SQL Server asignando más memoria, potencia de CPU y espacio en disco sin tener que reinstalar el software	- Dependen de otros servicios de Azure para su administración - Comunicaciones cifradas y firmadas mediante certificados - Características de Service Broker, Correo electrónico implica escoger instancia administrada - **Data Migration Assistant** comprueba la compatibilidad con un sistema local	- Se puede especificar una configuración sin servidor. En este caso las bases de datos que perteneces a varios suscriptores de Azure se comparten garantizando la privacidad. - Ofrece Advanced Threat Protection para la seguridad,evaluaciones de vulnerabilidad y corregir problemas de seguridad - Cifra datos en reposo y movimiento

Datos relacionales en Azure

Servicios de Azure para bases de datos de código abierto

Azure Database for MySQL	Azure Database for MariaDB	Azure Database for PostgreSQL
Tipos de servicio en la nube		
PaaS	PaaS	PaaS
Características		
- Alta disponibilidad - Copias de seguridad automática con restauración a un momento dado - Ofrece seguridad de conexión para aplicar reglas de firewall - Opcionalmente conexiones SSL - Muchos parámetros como modos de bloqueo, N° máximo de conexiones y tiempos de espera. - Sistema de base de datos global - Protección de datos en reposo y en movimiento - Funcionalidades de supervisión para agregar alertas y ver métricas y registros	- Alta disponibilidad - Copias de seguridad automática con restauración a un momento dado - Ofrece seguridad y cumplimiento de nivel empresarial (Similar a Azure Database for MySQL)	- Proporciona las mismas ventajas que MySQL en disponibilidad, rendimiento, escalado, seguridad y administración. - No están disponibles las extensiones para realizar tareas especializadas como escribir procedimiento almacenados en varios lenguajes de programación distintos de pgsql e interactuar con el S.O. - Tiene dos opciones de implementaciones: Servidor único e Hiperescala

Azure Database for PostgreSQL

Servidor único

- Ventajas similares a Azure Database for MySQL. Tres planes de tarifa : Básica, Uso general y Optimizado para memoria

Hiperescala (Citus)

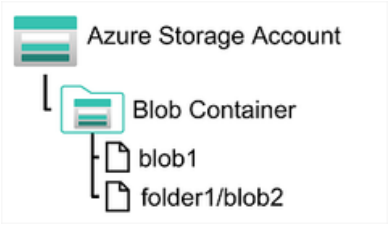
- Opción de implementación que escala las consultas entre varios nodos de servidor para admitir grandes cargas de base de datos.
- Una base de datos se divide en nodos y los datos se dividen en fragmentos según el valor de una clave de partición
- Ideal para bases de datos más grandes de PostgreSQL

Ventajas del servicio

- Mecanismos de conmutación por error y de detección de errores
- Registra información de las consultas que se ejecutan en las bases de datos del servidor y se guardan en una llamada *azure_sys*

Datos no-relacionales en Azure

Azure Blob Storage



- Permite almacenar grandes cantidades de datos no estructurados como blobs.
- Los blobs son una manera eficaz de almacenar archivos de datos
- Las aplicaciones pueden leerlos y escribirlos mediante la API de Azure Blob Storage
- Los blobs se almacenan en contenedores
- Puede controlar quién puede leer y escribir blobs dentro de un contenedor
- Dentro de un contenedor se puede organizar los blobs en una jerarquía de carpetas visuales

Tipos de blobs diferentes:

Blobs en bloques

- Un conjunto de bloques
- Cada bloque puede tener un tamaño de hasta 100 MB y un blob en bloques puede contener hasta 50 000. bloques con un tamaño máximo de más de 4,7 TB.
- El bloque es la cantidad más pequeña de datos que se puede leer o escribir como unidad individual.
- Usado para almacenar blobs discretos que cambian con poca frecuencia.

Blobs en páginas

- Colección de páginas de tamaño fijo de 512 bytes y puede contener en total hasta 8 TB de datos
- Optimizado para operaciones de lectura y escritura aleatorias
- Azure lo usa para implementar el almacenamiento de discos virtuales de las máquinas virtuales

Blobs en anexos

- Blob en bloques optimizado para admitir operaciones de anexión
- Solo se puede agregar bloques al final de un blob en anexos
- No se admite actualización o eliminación de bloques existentes
- Cada bloque puede tener hasta 4 MB y en total algo más de 195 GB

Tipos de niveles de acceso

Nivel de acceso frecuente

- Predeterminado. Los blobs se almacenan en medios de alto rendimiento

Nivel de acceso esporádico

- Rendimiento inferior e incurre en cargos de almacenamiento reducidos.
- Blobs a los que se accede con poca frecuencia.

Nivel de acceso archivo

- Proporciona el menor costo de almacenamiento, pero una mayor latencia.
- Se almacenan en un estado sin conexión
- Para recuperar un blob desde este nivel ,deberá cambiar el nivel de acceso a esporádico o frecuente.
- Solo puede leer el blob una vez que se ha completado el proceso de rehidratación

Directivas de administración de ciclo de vida para los blobs en una cuenta de almacenamiento

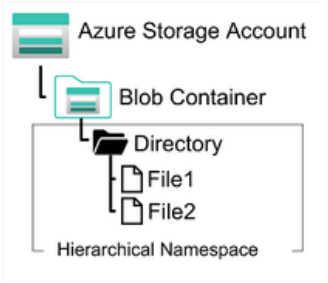
- Puede trasladar automáticamente un blob de acceso frecuente a uno esporádico o archivo, a medida que pasa el tiempo
- Puede organizarse para eliminar blobs obsoletos

Datos no-relacionales en Azure

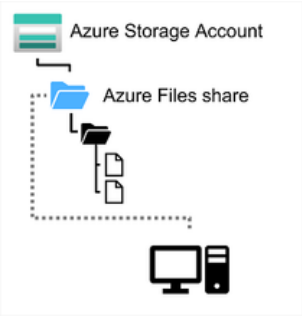
Azure Data Lake Storage Gen 2

- (Gen 1). Es un servicio independiente para el almacenamiento jerárquico de los datos de lagos de datos analíticos.
- Gen 2 es una versión más reciente de este servicio que se integra en Azure Storage
- Sistemas como Hadoop en Azure HDInsight, Azure Databricks y Azure Synapse Analytics pueden montar un sistema de archivos distribuido hospedado en Azure Data Lake Store Gen 2.

Para crear un sistema de archivos de Azure Data Lake Store Gen 2, se debe habilitar la opción "**Espacio de nombres jerárquico**" de una cuenta de Azure Storage. Este proceso es unidireccional, no podrá revertirlo



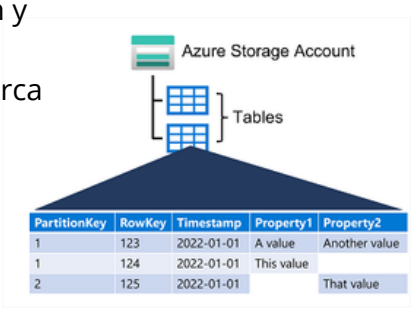
Azure Files



- Manera de crear recursos compartidos de red basados en la nube, similar a los que se encuentran en organizaciones locales para que los documentos y otros archivos estén a disposición de varios usuarios.
- Permite compartir hasta 100 TB de datos
- Estos datos se pueden distribuir en cualquier número de recursos
- Tamaño máximo de un solo archivo es de 1 TB, pero puede establecer cuotas de limite
- Admite hasta 2000 conexiones simultáneas por cada archivo compartido
- usar Azure File Sync para sincronizar copias almacenadas locamente en caché de archivos compartidos
- Dos niveles : Estándar (Disco duro) y Premium (Estado sólido)
- Dos protocolos de uso compartido de archivos de red:
  - Bloque de mensajes del servidor (SMB).
  - Network File System (NFS) . Requiere nivel Premium y una red virtual desde la que se controla el acceso al recurso compartido.

Azure Tables

- Solución de almacenamiento NoSQL
- Usa tablas que contienen elementos de datos de clave-valor.
- Cada elemento se representa mediante una fila que contiene columnas para los campos de datos que deben almacenarse
- Todas las filas deben tener una clave única (compuesta de una clave de partición y una clave de fila)
- La modificación de datos de una tabla se realiza también en una columna de marca de tiempo que registra la fecha y hora en que se realizó.
- Las columnas de cada fila pueden variar
- No hay conceptos de claves externas, relaciones, etc.



Azure Table Storage divide una tabla en particiones

- Una partición agrupa filas relacionadas según una propiedad común o clave de partición
- Las filas que comparten partición se almacenan juntas y una tabla puede contener cualquier número de particiones
- Al buscar datos, se puede incluir una clave de partición en los críterios de búsqueda

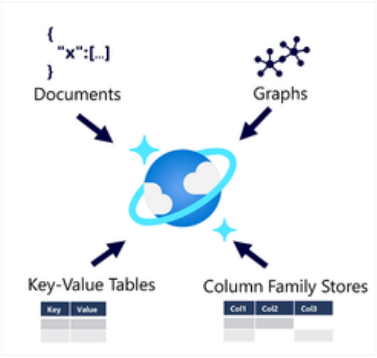
Datos no-relacionales en Azure

Azure Cosmos DB

- Servicio de baes de datos en la nube altamente escalable para datos NoSQL
- Admite varias interfaces de programación de aplicaciones (API) para usar tipos comunes de almacén de datos.
- La estructura de datos interna se abstrae
- Usa índices y particiones para proporcionar un rendimiento rápido de lectura y escritura
- Puede habilitar escrituras en varias regiones agregando las regiones de Azure que prefiera a su cuenta de Cosmos DB para que los usuarios puedan trabajar con datos en su réplica local.
- Asigna automáticamente espacio para las particiones en un contenedor y cada partición no crece más de 10 GB así como la creación de índices.

Casos de usos de Azure Cosmos DB

- **IoT y telemática**
- **Comercio y marketing**
- **Juegos**
- **Aplicaciones web y para móviles**



API de Azure Cosmos DB

API CORE (SQL)

- API nativa de Cosmos DB
- **Administra** los datos en formato de **documento JSON**
- Usa sintaxis SQL para trabajar con los datos y sus resultados no son tablas, sino documentos JSON

```SQL
SELECT *
FROM customers c
WHERE c.id = "joe@litware.com"
```

MongoDB API

- Es una base de datos de código abierto en la que los datos se almacenan en formato JSON binario (BSON)
- Usa la sintaxis MongoDB Query Languaje (MQL) compacta y orientada a objetos en la que se usan objetos para llamar a métodos
- Los resultados constan de documentos JSON

```JavaScript
db.products.find({id: 123})
```

```JSON
{
   "id": 123,
   "name": "Hammer",
   "price": 2.99}
}
```

Table API

- Trabaja con datos en tablas de clave-valor similar a Azure Table Storage
- Ofrece mayor escalabilidad y rendimiento que Azure Table Storage

PartitionKey	RowKey	Nombre	Email
1	123	Joe Jones	joe@litware.com
1	124	Samir Nadoy	samir@northwind.com

```text
https://endpoint/Customers(PartitionKey='1',RowKey='124')
```

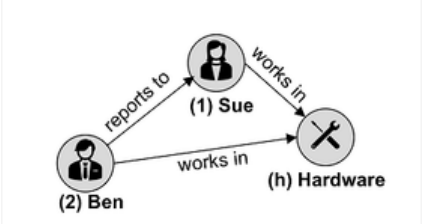
Cassandra API

- Compatible con Apache Cassandra, que usa una estructura de almacenamiento de familia de columnas
- Admite una sintaxis basada en SQL

```SQL
SELECT * FROM Employees WHERE ID = 2
```

Gremlin API

- Se usa con datos en una estructura de grafos
- Las entidades se definen como vértices que forman nodos en el gráfico conectado
- Los nodos se conectan mediante bordes que representan relaciones



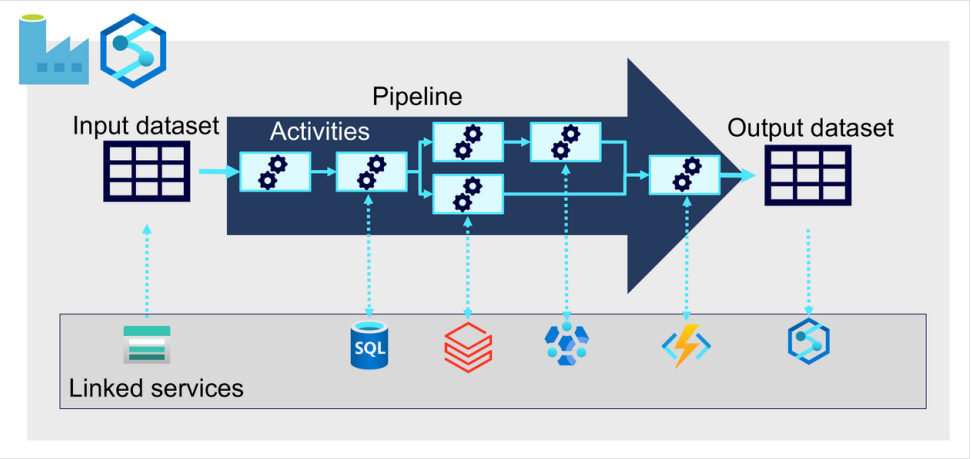

```
g.addV('employee').property('id', '3').property('firstName', 'Alice')
g.V('3').addE('reports to').to(g.V('1'))
```

Análisis de datos en Azure

Canalizaciones de ingesta de datos

Explicación de cómo se ingieren los datos en un almacén de datos analíticos de uno o varios orígenes

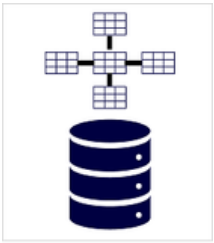
- La ingesta de datos a gran escala se implementa mejor mediante creación de canalizaciones que organicen procesos de ETL.
- Azure Data Factory puede crear y ejecutar canalizaciones o puede usar el mismo motor en Azure Synapse Analytics mediante sus Pipelines
Las canalizaciones constan de una o varias actividades que operan en los datos.



- Las actividades se pueden definir como un flujo de datos que manipula hasta que se genera un conjunto de datos de salida.
- Las canalizaciones utilizan servicios vinculados para cargar y procesar datos, y esto le permite usar la tecnología adecuada para cada paso del flujo de trabajo.

Almacenes de datos analíticos

Almacenamientos de datos



- Es una base de datos relacional en la que los datos se almacenan en un esquema optimizado para el análisis de datos (lectura)
- Normalmente los esquemas se basan en almacenar los datos en tablas de hechos centrales y dimensiones (esquema de estrella o esquema de copo de nieve)

Lagos de datos

- Es un almacén de archivos situado en un sistema de archivos distribuidos para el acceso a datos de alto rendimiento
- Se usan tecnologías como Spark o Hadoop para procesar consultas en los archivos almacenados
- Excelentes para admitir una combinación de datos estructurados, semiestructurados y no estructurados

Enfoques híbridos

- Combina características de lagos y almacenamientos de datos en una base de datos de lago o un lago de almacenamiento de datos.
- Los datos sin procesar se almacenan como archivos en un lago de datos y una capa de almacenamiento realacional abstrae los archivos y los expone como tablas que se pueden consultar mediante SQL (Los grupos de SQL de Synapse Analytics incluyen PolyBase que permiten definir tablas externas basadas en archivos de un lago de datos y consultarlas mediante SQL)

Análisis de datos en Azure

Fundamentos de la analítica en tiempo real

Procesamiento de flujos y por lotes

El procesamiento es la conversión de datos sin procesar en información significativa. Existen dos métodos generales:

- Procesamiento por lotes, en el que se recopilan y almacenan varios registros de datos antes de procesarse juntos en una sola operación.
- Procesamiento de flujos, en el que un origen de datos se supervisa y procesa constantemente en tiempo real a medida que se producen nuevos eventos de datos.

Ventajas de lotes

Se pueden programar y procesar grandes volúmenes de datos en un momento especificado.

Desventajas de lotes

El tiempo de retardo entre la ingesta de los datos y la obtención de los resultados.
Todos los datos de entrada de un trabajo por lotes deben estar listos para poder procesar un lote.

Ámbito de los datos

El procesamiento por lotes puede procesar todos los datos del conjunto de datos. Normalmente, el procesamiento en streaming solo tiene acceso a los datos recibidos más recientemente recibidos.

Rendimiento

La latencia del procesamiento por lotes suele ser de unas horas. Normalmente, el procesamiento en streaming se produce inmediatamente.

Tamaño de los datos

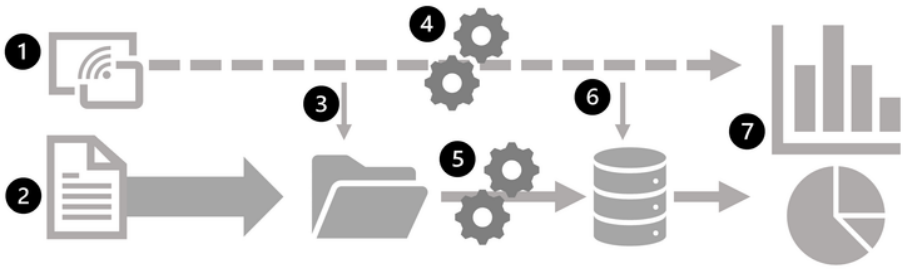
El procesamiento por lotes es adecuado para administrar grandes conjuntos de datos de forma eficaz. El procesamiento en streaming está diseñado para registros individuales o microlotes que constan de pocos registros.

Análisis

Normalmente se usa el procesamiento por lotes para realizar análisis complejos. El procesamiento en streaming se usa para funciones de respuesta simples, agregaciones o cálculos.

Combinación del procesamiento por lotes y por flujos

- Permite el análisis de datos históricos y en tiempo real
- Soluciones de procesamiento de flujos capturen datos en tiempo real, los procesen al tiempo que se conservan los resultados procesados en un almacén de datos



1. Los eventos de datos se capturan en tiempo real.
2. Los datos de otros orígenes se ingieren en un almacén de datos para el procesamiento por lotes.
3. Los datos de flujos se escriben en el almacén de datos

4. Procesamiento de flujos para preparar los datos de flujos para el análisis o visualización en tiempo real.
5. Los datos que no son de flujos se procesan por lotes periódicamente para prepararlos para el análisis.
6. Los resultados del procesamiento de flujos también se pueden conservar en el almacén de datos analíticos para admitir el análisis histórico.
7. Las herramientas analíticas y de visualización se usan para presentar los datos históricos y en tiempo real.

Análisis de datos en Azure

Procesamiento de flujos en Azure

Tecnlogías de procesamiento de flujos

Azure Stream Analytics

PaaS. Define trabajos de flujos que ingieren datos de un origen de flujos, aplican una consulta perpetua y escriben los resultados en una salida.

Spark Structured Streaming

Biblioteca de código abierto que le permite desarrollar soluciones de flujos complejos en servicios basados en Apache Spark

Orígenes para el procesamiento de flujos

Azure Event Hubs

Servicio de ingesta de datos que puede usar para administrar colas de datos de eventos.

Azure IoT Hub

Servicio de ingesta optimizado para administrar datos de eventos de IoT.

Azure Data Lake Store Gen 2

Servicio de almacenamiento que se puede usar como origen de datos de flujos

Apache Kafka

Solución de ingesta de datos de código abierto que se usa a menudo junto con Apache Spark

Receptores para el procesamiento de flujos

Azure Event Hubs

Para poner en cola los datos procesados

Azure Data Lake Store Gen 2 o Blob Storage

Conservar los resultados procesados como un archivo

Azure SQL Database, Synapse Analytics o Databricks

Conservar resultados procesados en una tabla de base de datos

Microsoft Power BI

Generar visualizaciones de datos en tiempo real

Delta Lake

Capa de almacenamiento de código abierto que agrega compatibilidad con la coherencia transaccional, cumplimiento del esquema y otras características comunes de almacenamiento de datos a Data Lake Storage.
- Unifica el almacenamiento para datos por lotes y de flujos
- Cuando se usa para el procesamiento de flujos, una tabla de Delta Lake se puede usar como un origen de flujos para las consultas en datos en tiempo real o como un receptor en el que se escribe un flujo de datos.

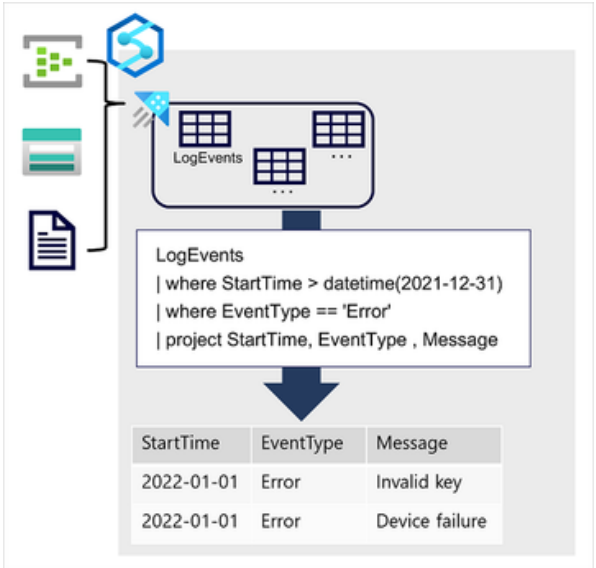
Azure Data Explorer

- Servicio independiente para analizar datos de manera eficiente.
- Los datos se introducen a través de uno o más conectores que permiten ingerir datos de fuentes estáticas o de transmisión.
- Los datos ingeridos se almacenan en tablas en una base de datos de Data Explorer, donde se permite consultas.

Lenguaje de consulta de Kusto (KQL)

Lenguaje optimizado para un rendimiento de lectura rápido, particularmente con datos de telemtría que incluyen un atributo de marca de tiempo

```
LogEvents
| where StartTime > datetime(2021-12-31)
| where EventType == 'Error'
| project StartTime, EventType , Message
```

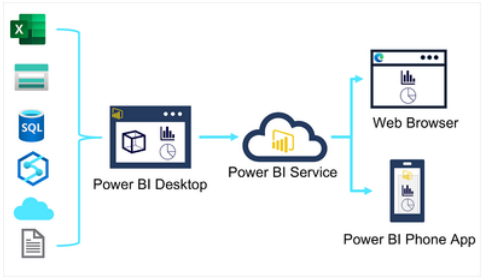


Análisis de datos en Azure

Visualización de datos en Azure

Herramientas y flujo de trabajo de Power BI

Existe muchas herramientas de visualización de datos y PowerBI es una solución integrada que puede admitir modelado de datos completo, informes interactivos y el uso compartido seguro.
El flujo de trabajo típico en PowerBI para crear una solución de visualización de datos:
- Power BI Desktop (Permite importar datos de una amplia gama de orígenes, combinar y organizar los datos en un modelo de datos de análisis)
- Crear informes que contengan visualizaciones interactivas de los datos
- Puede publicarlos en el servicio Power BI
- Los usuarios pueden consumir informes , paneles y aplicaciones mediante explorador web o dispositivo móvil.



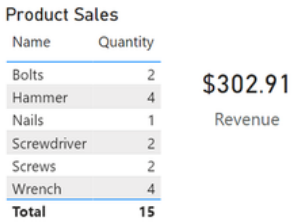
Jerarquías de atributos

- Aspecto sobre modelos analíticos que permiten rastrear agrupando datos o explorar en profundidad rápidamente

Consideraciones para visualizar datos

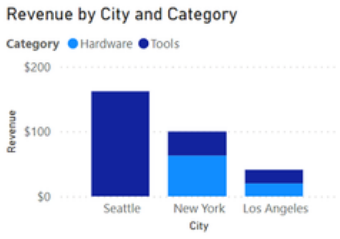
Tablas y texto

Útiles para mostrar numerosos valores relacionados



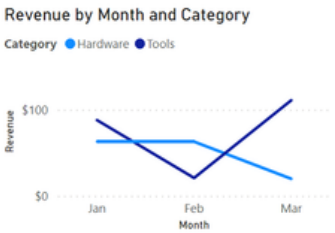
Gráficos de barras y columnas

Buena manera de comparar visualmente valores numéricos para categorías discretas



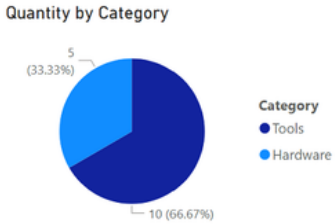
Gráficos de líneas

- Comparar valores clasificados
- Útiles cuando se examina tendencias a lo largo de una dimensión (a menudo el tiempo)



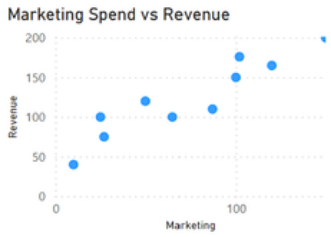
Gráficos circulares

Comparar visualmente los valores como proporciones de un total



Gráficos de dispersión

Comparar dos medidas numéricas e identificar una relación o correlación entre ellas



Mapas

Comparar visualmente valores de diferentes áreas geográficas o ubicaciones