

DP-900 Data Fundamentals

martes, 21 de febrero de 2023 14:12

CLASE 1 – 20_Febrero_23:

Sesión 1: Conceptos básicos de datos:

1. ¿Qué es un dato?

- Colección (una lista o conjunto de datos) de hechos, números, descripciones, objetos, almacenados de forma estructurada (tablas), semiestructurada (JSON) y no estructurada (audio, imágenes, fotos, etc.)

1.1. Clasificación:

1.1.1. Datos estructurados:

- Tablas ordenadas con llaves principales y foráneas para conectar con otras tablas. Bases de datos relacionales - SQL. (MySQL, postgres)

Procesamiento Transaccional en Línea (OLTP)	Procesamiento Analítico en Línea (OLAP)
Bases de datos que me permiten hacer transacciones (agregar, actualizar, eliminar) en el menor tiempo posible. Bases de datos enfocados en el día a día. Se alimenta en vivo.	Bases de datos enfocadas en analizar. Ofrecen métricas de un periodo de tiempo mayor a un día. Me permite contestar preguntas que me hacen al negocio. Tiene una fecha de corte en el tiempo.

1.1.2. Datos semiestructurados:

- Datos con estructura flexible, generalmente archivos tipo JSON con estructura llave-valor. Bases de datos no relacionales - No-SQL. (MongoDB, Casandra, postgres)

- **1.1.3. Datos no estructurados:**

- Datos que no se ajustan a una estructura específica tipo audio, imágenes, fotos, videos, etc.

1.2. Sistema de Análisis:

os de
etc.).

datos

)
n
n día.
e un
o.

Bases de

Preguntas iniciales:

1. **¿En dónde se encuentran mis datos?**

Ingesta de los datos: proceso de captura de datos desde diferentes orígenes para llevarlos a un sistema de almacenamiento (Batch/Streaming)

Datos por lotes Basch	Datos de Streaming
Registra dato por dato de cada dispositivo y lo va guardando. Se hace una sincronización periódica.	Cada que toma un dato se envía. Se necesita menor procesamiento y más capacidad.

2. **¿Qué tipo de datos voy a recolectar?**

Almacenamiento de datos: Dependiendo del tipo de datos, defino cómo guardarlos (relacional, no relacional, etc.)

3. **¿Cómo los vamos a procesar?**

Procesamiento de datos: se define el tipo de tratamiento para responder a las preguntas del negocio.

4. **¿Cómo voy a mostrar las conclusiones?**

Visualización de datos: se muestran los resultados del análisis de los datos en gráficas.

Lección 1: Comprobación de conocimientos



¿Cómo se organizan los datos en una tabla relacional?

- Filas y columnas
- Encabezado y pie de página
- Páginas y párrafos



¿Cuál de las siguientes opciones es un ejemplo de datos no estructurados?

- Una tabla de empleados con columnas con id. de empleado, nombre y designación de empleado
- Archivos de audio y video
- Una tabla dentro de la base de datos de SQL Server



¿Cuál de las siguientes opciones es un ejemplo del conjunto de datos de streaming?

- Datos de las fuentes del sensor
- Datos de ventas del último mes
- Lista de empleados que trabajan para una compañía

Sesión 2: Explorar roles y responsabilidades en el mundo de los datos:

1. **Roles y funciones:**

stema de

nás

o

ocio.

Administrador de base de datos:	Ingeniero de datos:	Analista de datos
<ul style="list-style-type: none"> • Administración de base de datos • Implementa la seguridad de los datos • Copias de seguridad • Acceso a usuarios • Supervisa el rendimiento • Trabaja con el sistema que va a almacenar los datos, pero con ellos puntualmente. 	<ul style="list-style-type: none"> • Procesos y canalizaciones de datos • Almacenamiento de ingesta de datos • Prepara datos para el análisis • Prepara datos para el procesamiento analítico • No analiza los datos (va desde la ingesta, hasta el procesamiento) 	<ul style="list-style-type: none"> • Proporciona conclusiones sobre los datos • Informes visuales • Modelado de datos para análisis • Combina datos para visualización y análisis • No procesa datos, los toma procesados y analiza.

2. Herramientas comunes:

2.1. Administrador de la base de datos:

Azure Data Studio	SQL Server Management Studio	Azure Portal/CLI
<ul style="list-style-type: none"> • Interfaz gráfica para administrar servicios de datos locales y basados en la nube • Se ejecuta en Windows, macOS, Linux 	<ul style="list-style-type: none"> • Interfaz gráfica para administrar servicios de datos locales y basados en la nube • Se ejecuta en Windows • Herramienta integral de administración de bases de datos 	<ul style="list-style-type: none"> • Herramientas para la administración y el aprovisionamiento de Azure Services. • Ejecución manual y automatizada de scripts usando Azure Resource Manager o interfaz scripting de la línea de comandos.

2.2. Ingeniero de Datos:

Azure Synapse Studio	SQL Server Management Studio	Azure Portal/CLI
<ul style="list-style-type: none"> • Azure Portal integrado para administrar Azure Synapse. • Ingesta de datos (Azure Data Factory) • Administración de recursos de Azure Synapse (grupos de SQL/grupo de Spark) 	<ul style="list-style-type: none"> • Interfaz gráfica para administrar servicios de datos locales y basados en la nube • Se ejecuta en Windows • Herramienta integral de administración de bases de datos 	<ul style="list-style-type: none"> • Herramientas para la administración y el aprovisionamiento de recursos de Azure • Ejecución manual y automatizada de scripts usando Azure Resource Manager o interfaz de scripting de la línea de comandos

Data
ilizada
z de

rsos
ando
nea

2.3. analista de Datos:

Power BI Desktop	Portal de Power BI/Servicio Power BI	Power BI Report Builder
<ul style="list-style-type: none">• Herramienta de visualización de datos• Modelar y visualizar datos• Administración de recursos de Azure Synapse (grupos de SQL/grupo de Spark)	<ul style="list-style-type: none">• Crear y administrar informes de Power BI• Crear paneles de Power BI• Compartir informes/conjuntos de datos	<ul style="list-style-type: none">• Herramienta de visualización de datos para informes paginados• Modelar y visualizar informes paginados

Lección 2: Comprobación de conocimientos



¿Cuál de las siguientes tareas es un rol de un administrador de base de datos?

- Copias de seguridad y restauración de bases de datos
 Crear paneles e informes
 Identificar problemas de calidad de datos



¿Cuál de las siguientes herramientas es para la visualización y generación de informes?

- SQL Server Management Studio
 Power BI
 SQL



¿Cuál de los siguientes roles no es una función de datos?

- Administrador de sistemas
 Analista de datos
 Administrador de base de datos

Sesión 3: Describir conceptos de datos relacionales

1. **Tablas:** Colección de datos que se almacenan en un lugar específico.
 - Estructura de almacenamiento de datos
 - La tabla consta de filas (entidades) y columnas
 - Todas las filas tienen el mismo número de columnas
 - Cada columna está definida por un tipo de dato
2. **Normalización:** conjunto de reglas para segmentar la información. Los datos se normalizan para:
 - Reducir el almacenamiento
 - Evitar la duplicación de datos

de
os
s

- Mejorar la calidad de los datos

2.1. Características de un esquema de base de datos normalizado:

- Las llaves primarias y externas/foráneas se utilizan para definir relaciones
- No existe duplicación de datos (excepto los valores de clave en 3a forma normal (3NF)).
- Los datos se recuperan uniendo tablas en una consulta.

3. **Índice:** se usa para encontrar información un poco más rápido.

- Optimiza las consultas de búsqueda para una recuperación de datos más rápida
- Reduce la cantidad de páginas de datos que deben leerse para recuperar los datos en una instrucción SQL
- Los datos se recuperan uniendo tablas en una consulta.

4. **Vista:** es una tabla virtual que se basa en el conjunto de resultados de la consulta:

- Las vistas se crean para simplificar la consulta
- Combinan datos relacionales en una vista única de panel

Lección 3: Comprobación de conocimientos



¿Cuál de las siguientes afirmaciones es una característica de una base de datos relacional?

- Todos los datos deben almacenarse como cadenas de caracteres.
- Una fila en una tabla representa una sola entidad
- Diferentes filas en la misma tabla pueden contener diferentes columnas



¿Qué es un índice?

- Una estructura que le permite ubicar filas en una tabla rápidamente, usando un valor indexado
- Una tabla virtual basada en el conjunto de resultados de una consulta
- Una estructura que comprende filas y columnas que usa para almacenar datos

CLASE 2 – 22_Febrero_23:

Sesión 4. Explorar conceptos de datos no relacionales:

- Las **colecciones no relacionales** pueden tener:
 - Varias entidades en la misma colección o contenedor con campos diferentes
 - Un esquema diferente no tabular
 - Se suelen definir etiquetando cada campo con el nombre que representan

Identificar casos de uso de bases de datos no relacionales



IoT y telemática:

A menudo requieren ingerir grandes cantidades de datos en ráfagas frecuentes de actividad, los datos son semiestructurados o estructurados, a menudo requieren procesamiento en tiempo real



Comercio y marketing:

Escenarios comunes para datos distribuidos globalmente, almacenamiento de documentos



Juegos:

Estadísticas del juego, integración en redes sociales, marcadores, aplicaciones de baja latencia



Web y móvil:

Se suelen usar con análisis de clics en web y aplicaciones modernas que incluyen bots

- **Tipos de datos no relacionales:**

- La estructura de los datos se define en los mismos datos por medio de campos. Los tipos de formato/archivo incluyen:
 - JSON, AVRO, ORC, Parquet

¿Qué es NoSQL?

Término suelto para describir los **modelos de almacenamiento de datos No Relacional**. Ejemplos de Bases de Datos:

- **Almacenes Clave-valor:**

- Se utiliza como almacenamiento de elementos IoT. Permiten guardar datos muy rápido, fáciles de consultas rápido.
- Se almacena aglomerando información. No se deben hacer actualizaciones, se demora mucho el busca información por sus claves.
- Un almacén clave-valor asocia cada valor de datos con una clave única. La mayoría de los almacenes clave-valor solo admiten operaciones simples de consulta, inserción y eliminación.
- Son menos adecuados si necesita consultar datos en diferentes almacenes de clave-valor, los almacenes de clave-valor tampoco están optimizados para realizar consultas por valor.

Key	Value
AAAAAA	1101001111010100110101111...

ses de

cil y hacer

cho. Se

lmacenes

los

AABAB	100110000101100110101110...
DFA766	0000000000101010110101010...
FABCC4	1110110110101010100101101...

- **Basado en documentos:**

- Almacena una colección de documentos, donde cada documento consta de datos y campos nombre. Los datos pueden ser valores simples o elementos complejos como listas y colecciones secundarias. Los documentos se recuperan mediante claves únicas.

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }
1002	{ "CustomerID": 220, "OrderItems": [{ "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" }

- **Familias de columnas:**

- Es muy similar a una base de datos relacional, el poder deal de una base de datos de familias de columnas radica en su enfoque desnormalizado para estructurar datos dispersos.
- Los datos se almacenan en tablas que constan de una columna clave y una o más familias de columnas (grupos de información tipo categorías). Enfoque desnormalizado para estructurar datos dispersos.

s con
ones

as de

e
ar datos

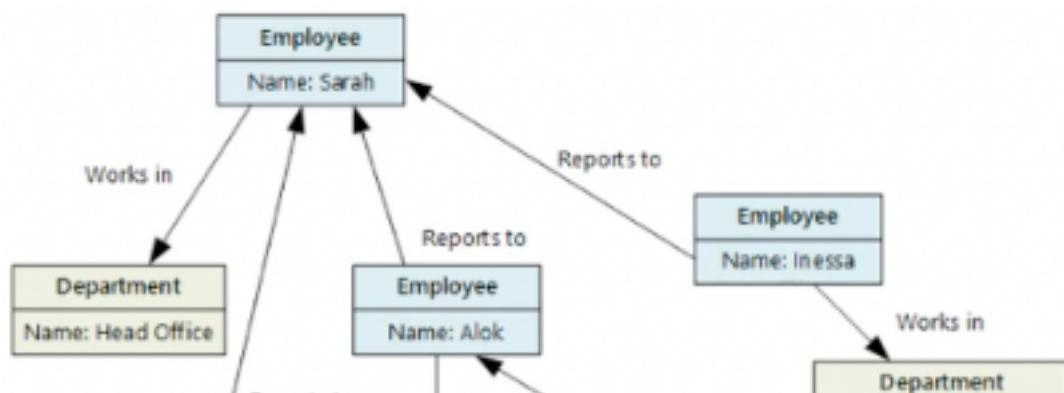
CustomerID	Column Family: Identity
001	First name: Mu Bae Last name: Min
002	First name: Francisco Last name: Vila Nova Suffix: Jr.
003	First name: Lena Last name: Adamczyz Title: Dr.

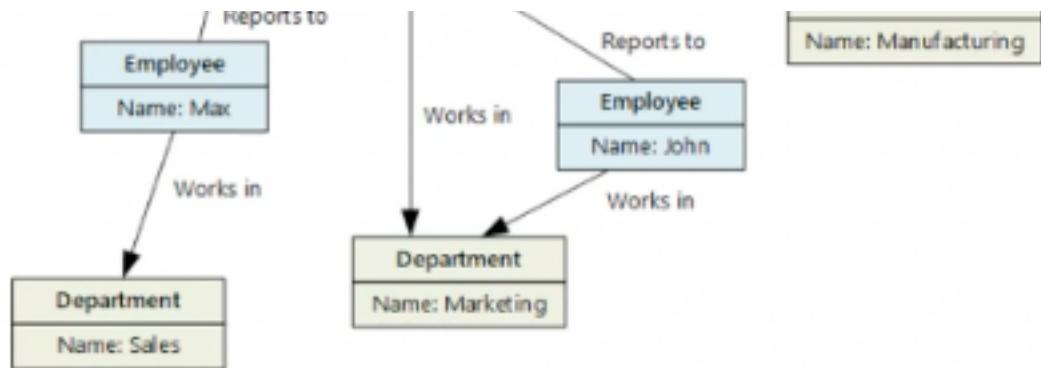
CustomerID	Column Family: Contact Info
001	Phone number: 555-0100 Email: someone@example.com
002	Email: vilanova@contoso.com
003	Phone number: 555-0120

Row Key	Column Families			
	CustomerID	CustomerInfo		AddressInfo
1		CustomerInfo:Title Mr CustomerInfo:FirstName Mark CustomerInfo:LastName Hanson		AddressInfo:StreetAddress 999 500th Ave AddressInfo:City Bellevue AddressInfo:State WA AddressInfo:ZipCode 12345
2		CustomerInfo:Title Ms CustomerInfo:FirstName Lisa CustomerInfo:LastName Andrews		AddressInfo:StreetAddress 888 W. Front St AddressInfo:City Boise AddressInfo:State ID AddressInfo:ZipCode 54321
3		CustomerInfo:Title Mr CustomerInfo:FirstName Walter CustomerInfo:LastName Harp		AddressInfo:StreetAddress 999 500th Ave AddressInfo:City Bellevue AddressInfo:State WA AddressInfo:ZipCode 12345

- Grafos:**

- Almacenes de entidades centradas en relaciones.
- Permite que las aplicaciones realicen consultas atravesando una red de nodos y bordes.





© Copyright Microsoft Corporation. All rights reserved.

Lección 4: Comprobación de conocimientos



¿Cuál de los siguientes servicios debería utilizar para implementar una base de datos no relacional?

- Azure Cosmos DB
- Azure SQL Database
- La API de Gremlin



¿Cuál de las siguientes es una característica de las bases de datos no relacionales?

- Las bases de datos no relacionales contienen tablas con registros planos de columna fija
- Las bases de datos no relacionales requieren el uso de técnicas de normalización de datos para reducir la duplicación de datos
- Las bases de datos no relacionales no tienen esquemas o tienen esquemas laxos



Está creando un sistema que supervisa la temperatura en un conjunto de bloques de oficinas y configura el aire acondicionado en cada sala de cada bloque para mantener una temperatura ambiente agradable. Su sistema tiene que administrar el aire acondicionado en varios miles de edificios repartidos por el país o la región, y cada edificio suele contener al menos 100 salas con aire acondicionado. ¿Qué tipo de almacenamiento NoSQL es el más apropiado para capturar los datos de temperatura y así permitir que se procesen rápidamente?

- Un almacén de valores clave
- Una base de datos de familias de columnas
- Escribir las temperaturas en un blob de Azure Blob Storage

Sesión 5: Procesamiento de Datos, EDA y Visualización:

- **Servicios en la nube:** entrega de servicios informáticos (almacenamiento, nube, IS, BDs, recursos de aplicaciones web, aplicaciones en IoT, satélites, etc.)
- **Redundancia:** opciones de disponibilidad
 - **Sets de disponibilidad:** protege contra fallas. Se tiene una réplica del datacenter propio en zona. Alto riesgo de daños por desastres naturales. (NC 99.95%)
 - **Zonas de disponibilidad:** Varios datacenters. protección de fallas en datacenters enteros. NC 99.999%

como

una

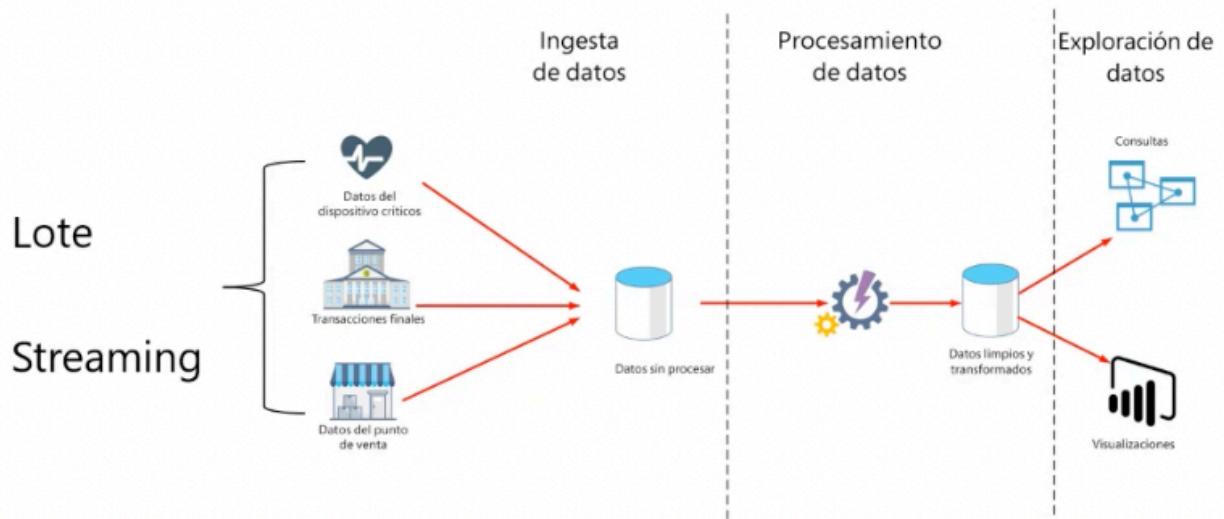
menor

riesgo de daños por desastres naturales. (NC 99.95%)

- **Recuperación multirregión ante desastres:** Regiones pares, sin límites de Data Residency. En caso de desastre, usa un datacenter en una región completamente opuesta.

1. Procesamiento de Datos:

- **Ingesta de Datos:**
- Es el proceso de obtener e importar datos para su uso inmediato o almacenamiento en una base de datos.



- **Procesamiento de Datos:**
- Qué tipo de herramientas/servicios vamos a usar para lograr la calidad de los datos (limpiar información basura, datos confidenciales, etc.).



- **Data Warehouse** (sistemas estructurados) se habla de big data, petabytes, permiten almacenar grandes cantidades enormes de información pero sólo en formatos que él conoce. sino se puede perder información.

En caso

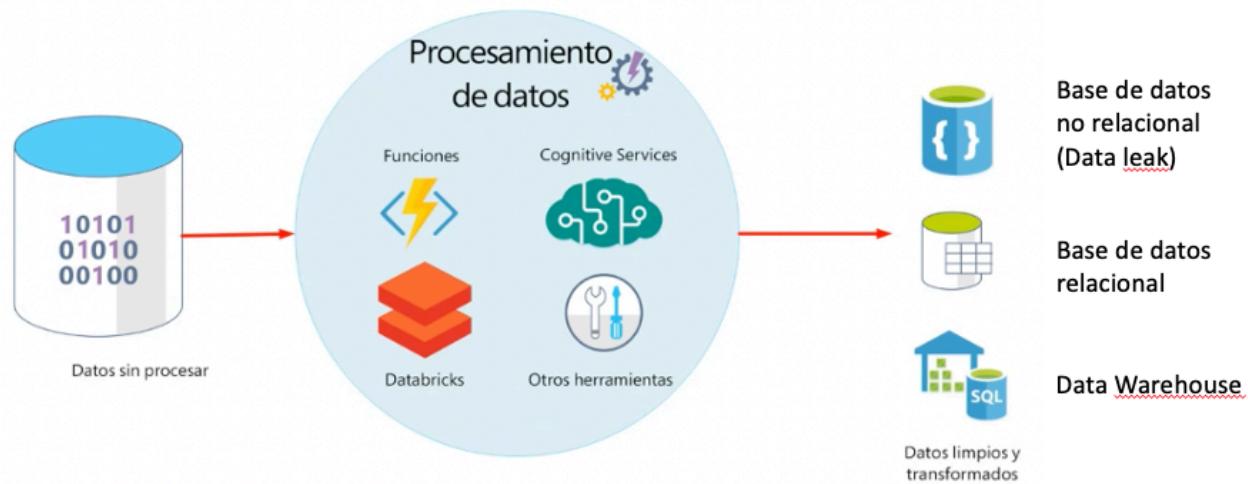
de datos.

mación

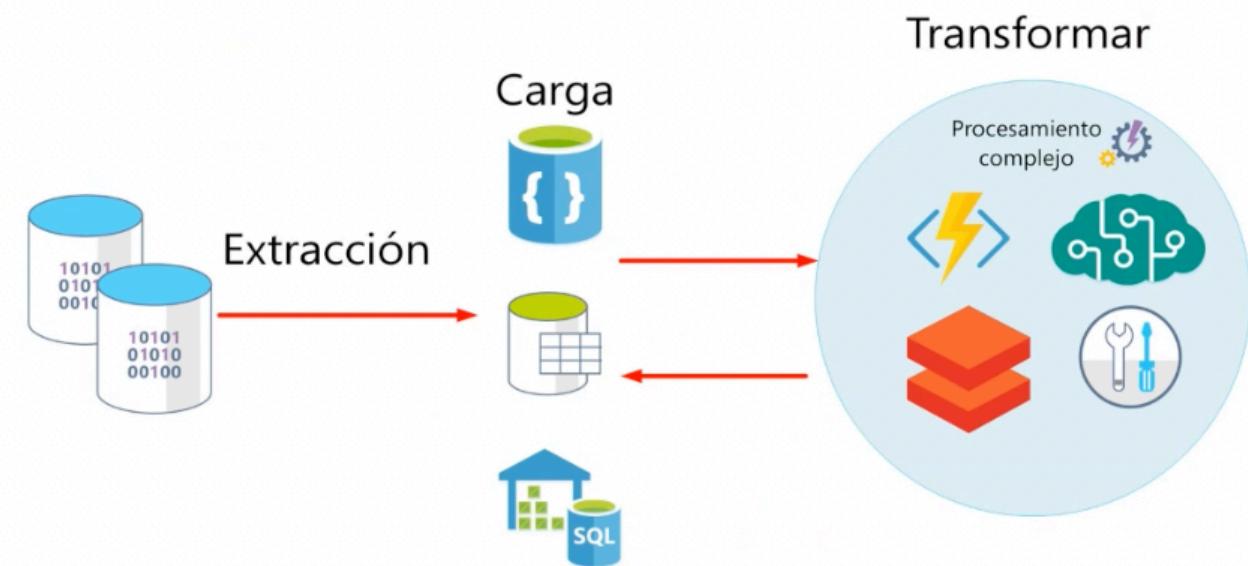
enar
erder

información (Importante ETL antes de guardar).

- **ETL (Extraction, Transforming and Loading):** Mecanismo de procesamiento de datos que descartan datos confidenciales, se extraen los demás datos, se transforman (filtrado y transformaciones de formatos) y se cargan a una base de datos.



- **Data Leak** (sistemas semiestructurados) se habla de big data, petabytes, permiten almacenar cantidades enormes de información sin procesarla previamente.
- **ELT (Extraction, Loading and Transforming):** Mecanismo de procesamiento de datos en el que el motor de procesamiento de datos puede adoptar un enfoque iterativo y a menudo es un tipo de procesamiento por lotes frecuente. Se guarda toda la información, se secciona por lotes y se procesa y se guarda en la base de datos final (SQL o NoSQL).



s. Se
básico y

ar

s. El
con
otes, se

2. Explorar el Análisis de Datos:

Análisis Descriptivo	Análisis Diagnóstico	Análisis Predictivo	Análisis Prescriptivo	Análisis Cognitivo
				
<ul style="list-style-type: none"> • ¿Qué sucedió/está sucediendo? • Trabaja en función de un histórico. 	<ul style="list-style-type: none"> • ¿Por qué sucedió/está sucediendo? • Investiga las causas de algo. 	<ul style="list-style-type: none"> • ¿Qué podría pasar? • Ayuda a saber qué puede pasar en el futuro, de acuerdo con el comportamiento histórico. 	<ul style="list-style-type: none"> • ¿Qué decisiones deberíamos tomar? • Decisiones para lograr un objetivo. 	<ul style="list-style-type: none"> • Datos no estructurados (imágenes, etc.) • Generar ideas sobre datos estructurados (patrones de IA).

3. Explorar la Visualización de Datos:

Power BI: Colección de software, servicios, aplicaciones y conectores. Conocimiento del negocio a través de los datos.



Power BI

álisis
nitivo



ados
s, videos,

nferencias
tos no
ados
, etc) con

a través



Lección 5: Comprobación de conocimientos



¿Qué es la ingesta de datos?

- El proceso de transformar datos sin procesar en modelos que contienen información significativa.
- Analizar datos en busca de anomalías
- Recopilar streaming de datos sin procesar de varios orígenes y almacenarlos



¿Cuál de los siguientes objetos visuales muestra los principales colaboradores a un resultado o valor seleccionado?

- Elementos influyentes clave
- Gráfico de columnas y barras
- Gráfico de matriz



¿Qué tipo de análisis responder preguntas sobre lo que sucedió en el pasado?

- Análisis descriptivo
- Análisis prescriptivo
- Análisis predictivo

27/febrero/2023:

Datos Relacionales en Azure:

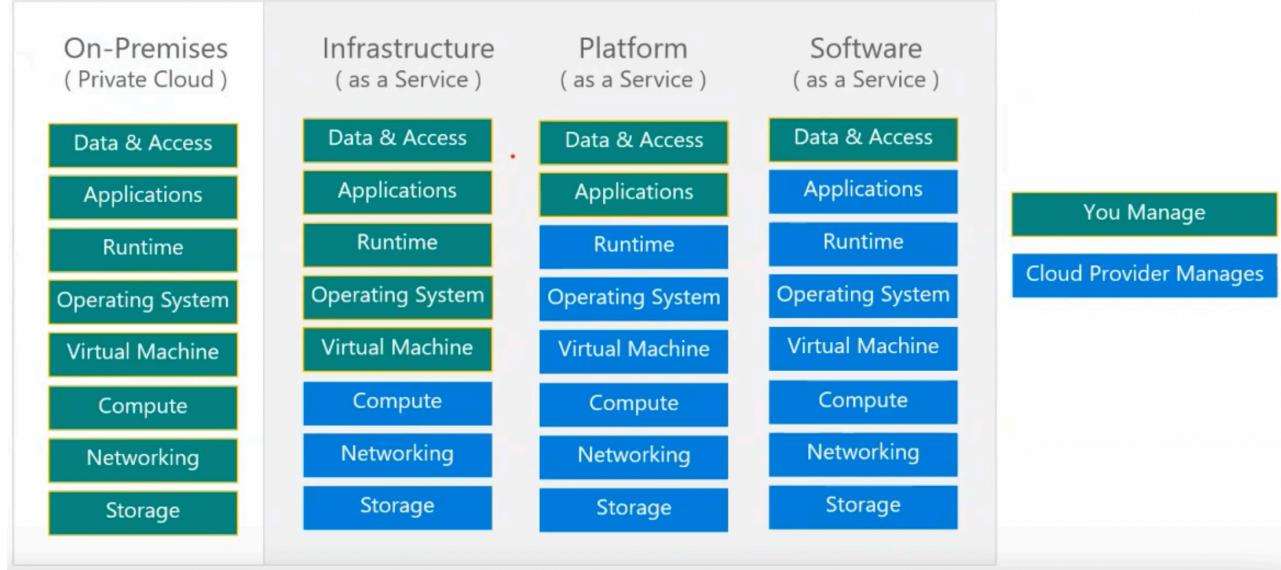
Modelo de responsabilidad compartida:

- **On-premises:** Servicios privados de una empresa, servidores físicos que se administran personalmente.
- **Infrastructure:** Algunos servicios son controlados por un proveedor. Cuenta con dos partes de responsabilidad: administración del proveedor de la nube (redes, almacenamiento y capacidades de cómputo) y administración propia (aplicaciones, sistema de operación, máquinas virtuales)

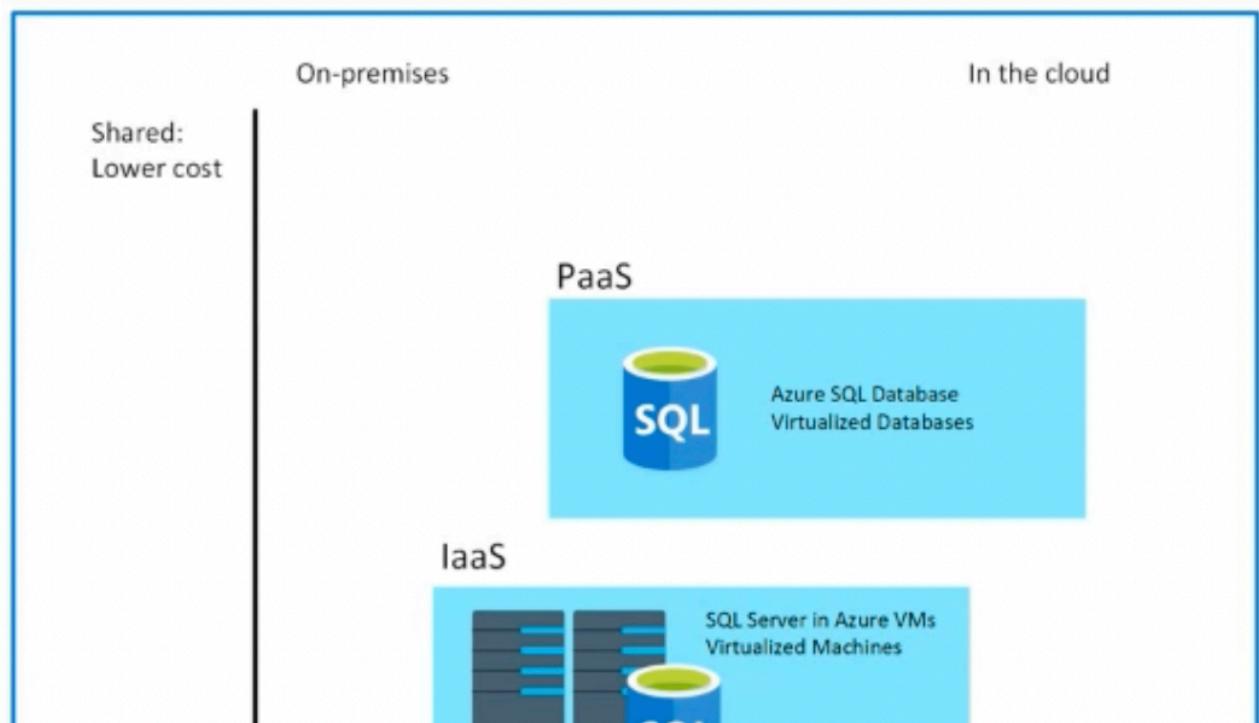
mente.

de

- **Platform:** Adicional al anterior, instala la máquina virtual, el sistema operativo y las actualizaciones del sistema. El usuario solo se encarga de las aplicaciones y los procesos de autenticación.
- **Software:** SAS, conjunto completo de aplicaciones (todo el paquete de 365), más todo el componente de Office. Lo único que se protege es el correo con usuario y contraseña y la información que se tiene almacenada.



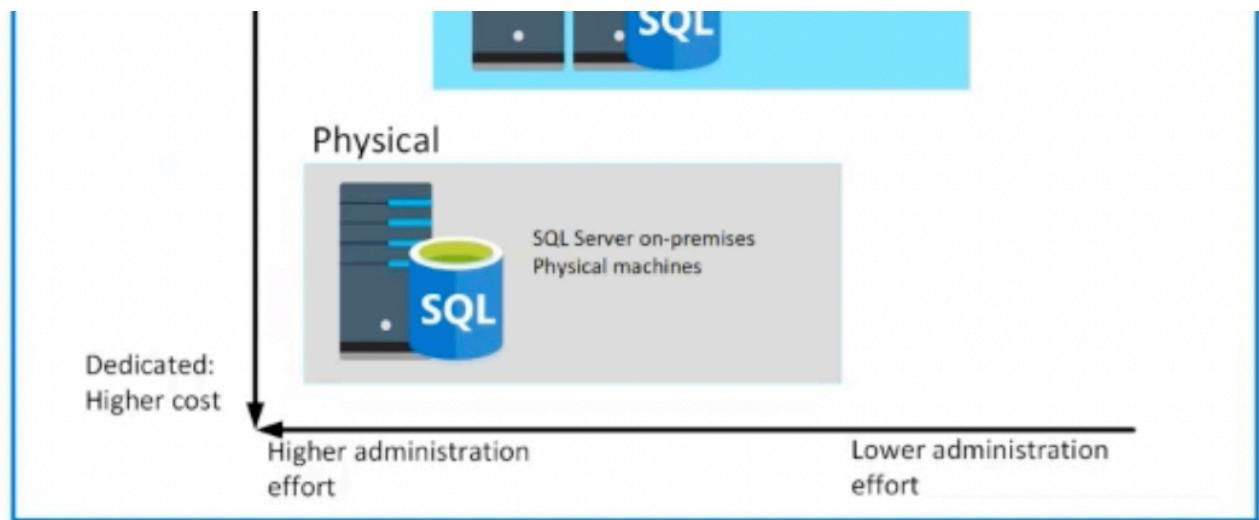
IaaS frente a PaaS:



s del

ente del

e



¿Qué es Azure Data Services?

- **SQL Server en Azure Virtual Machines: \$**
 - Lo mejor para rehospedaje y aplicaciones que requieres acceso y control en el nivel de sistema operativo.
 - Funciones de capacidad de administración automatizada y acceso de nivel de sistema operativo.
 - Infraestructura como servicio.
- **Instancia Administrada por Azure SQL: \$\$\$**
 - Lo mejor para modernizar aplicaciones existentes.
 - Ofrece alta compatibilidad con SQL server y soporte nativo de VNET
 - Plataforma como servicio.
- **Azure SQL Database: \$\$**
 - Lo mejor para crear nuevas aplicaciones en la nube
 - Proceso preaprovisionada o sin servidor y almacenamiento de hiperescala para cumplir con exigentes requisitos de la carga de trabajo
 - Plataforma como servicio.

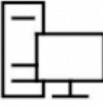
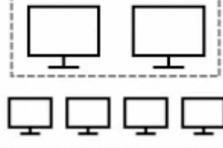
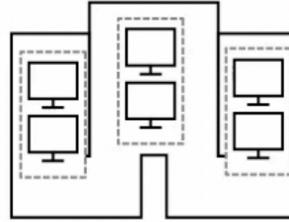
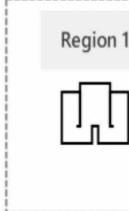
Determinar opciones de disponibilidad:

VM SLA 99.9% with Premium Storage	VM SLA 99.95%	VM SLA 99.99%	MULTI-REGIONAL RECOVERY
--------------------------------------	------------------	------------------	-------------------------

ema

ativo.

n los

				
SINGLE VM Easier lift and shift		AVAILABILITY SETS Protecting against failures within datacenters		AVAILABILITY ZONES Protection from entire datacenter failures
Servidor y capacidad de cómputo contratados. No tiene respaldo en vivo. Falla en caso de falta de electricidad o fallas naturales.		Tiene un respaldo en local que sincroniza la info del principal y entra a operar si se daña el primero. Está limitado a un datacenter. Buena resistencia a fallas.	Zonas a nivel regional especiales que tienen info sincronizada y se habilitan al dañarse la principal. Está limitado a una región. Buena resistencia ante desastres naturales.	Zonas sincronizadas al dañarse. Está repartida entre regiones resistentes a desastres.

Grupos Elásticos: Servidor y capacidad de cómputo se aprovisiona con buena capacidad. Se aprovecha porcentualmente entre las diferentes bases de datos. Permite crecer y disminuir según los requerimientos. Se tiene un solo servidor y se distribuye su capacidad de acuerdo a la necesidad de las BDs.

OLTP: necesitan más capacidad de cómputo que las OLAP. Salvo que estas últimas tengan modelos de machine learning.

Redundancia de almacenamiento: copia por si algo llega a pasar, no es una copia de consulta.

Ventajas de Azure Database for MySQL, PostgreSQL y MariaDB



Base de datos totalmente administrada de la comunidad:

Aprovechar un servicio totalmente administrado sin dejar de usar las herramientas y los lenguajes con los que está familiarizado



Alta disponibilidad incorporada para reducir el TCO

Asegúrese de que sus datos estén siempre disponibles sin la necesidad de costes adicionales



Rendimiento y escalado inteligentes:

Mejorar el rendimiento con inteligencia integrada y hasta 16 TB de almacenamiento y 20 000 IOPS



s
ection within
y Boundaries

especiales que
an y se habilitan
se la principal.
artido en varias
nes. La mejor
tencia ante
res naturales.

isiona
imientos.

s de



Seguridad y cumplimiento líderes en el sector:

Proteger sus datos con características de seguridad mejorada que incluyen Advanced Threat Protection

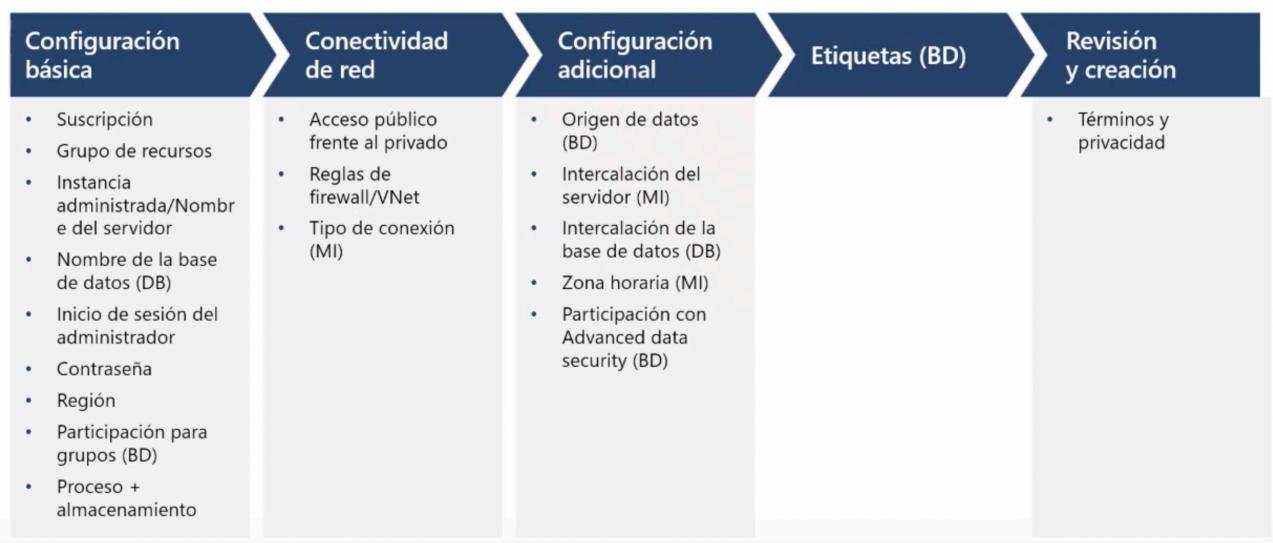


Integración con el ecosistema de Azure:

Crear aplicaciones más rápido con los servicios de Azure y proteger su innovación con Azure IP Advantage

Aprovisionar: decirle al sistema que necesitamos un servicio puntual y que el sistema cree el servicio con recursos.

Configurar servicios de datos relacionales



Control de acceso basado en roles de Azure (RBAC)

Todas las operaciones de Azure para Azure SQL se controlan a través de RBAC

Piense en esto como si fuesen permisos de seguridad fuera de la instancia administrada o de la base de datos

Entidad de seguridad y sistema basado en roles

El ámbito incluye la suscripción, el grupo de recursos y el recurso

Desacoplado de SQL Security (hoy)

Se aplica a las operaciones en Azure Portal y CLI

Permite la separación de tareas para la implementación, la administración y la utilización

Los bloqueos de Azure ayudan a proteger los recursos de la eliminación o del estado de solo lectura

Roles de Azure SQL integrados disponibles para reducir la necesidad de un propietario

Colaborador de BD SQL

Colaborador de instancia administrada de SQL

Administrador de seguridad de SQL

Colaborador de SQL Server

RBAC: control detallado de qué puede o no hacer cada usuario de la BD.

Réplicas de Lectura:

n los

BD de Azure: réplicas de lectura

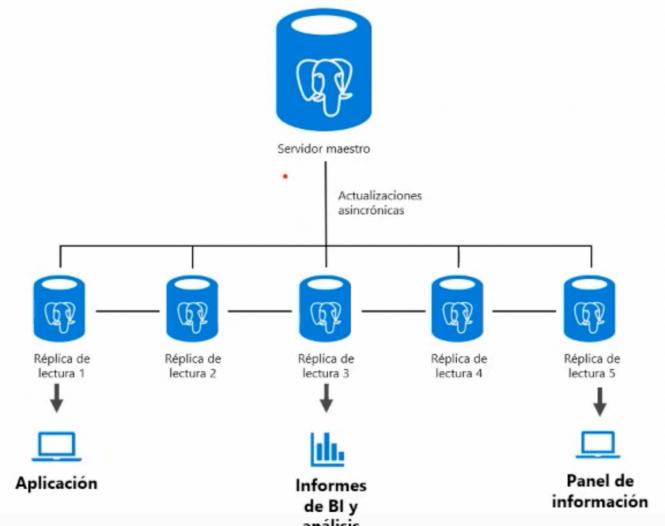
Mejorar el rendimiento y la escala de las cargas de trabajo

Escenarios en los que los retrasos en la sincronización de datos entre el maestro y las réplicas sean aceptables

Cree una réplica en una región de Azure diferente de la del maestro para tener un plan de recuperación ante desastres.

El almacenamiento de datos en los servidores de réplica crece automáticamente sin afectar las cargas de trabajo

Cree hasta cinco réplicas del servidor maestro



01/marzo/2023:

Consulta de datos relacionales:

- SQL es un lenguaje estándar usado en bases de datos relacionales.
- ANSI e ISO se encargan de mantener los estándares de SQL.

Tipos de Instrucciones de SQL:

DML Data Manipulation Language	DDL Data Definition Language	DCL Data Control Language
<ul style="list-style-type: none">• Lenguaje de <u>manipulación</u> de datos.• Se usa para consultar y manipular datos.• (SELECT, INSERT, UPDATE, DELETE) <p>*Ojo con el Update, siempre con Where de ID.</p>	<ul style="list-style-type: none">• Lenguaje de <u>definición</u> de datos.• Se utiliza para definir objetos de bases de datos.• (CREATE, ALTER, DROP)	<ul style="list-style-type: none">• Lenguaje de <u>control</u> de datos.• Se usa para <u>administrar</u> permisos de seguridad.• (GRANT, REVOKE, DENY)

6/marzo/2023:

Storage: Espacio de almacenamiento de Azure. Permite guardar cualquier tipo de dato. El límite lo define el usuario:

- **Blobs:** objetos binarios muy grandes: imágenes, videos, streaming.
- **Files:** tipo OneDrive, para compartir información
- **Table:** Datos semiestructurados tipo clave- valor
- **Q:** colas, conexiones con diferentes servicios – mensajería.

Azure Table Storage:

- Almacenes clave-valor.
- La clave se compone de dos elementos:
 - **Clave de partición** (identifica la partición que contiene la fila) y
 - **Clave de fila** (única para cada fila de la misma partición).

Azure Blob Storage:

Se pueden guardar por carpetas, como un OneDrive. Las carpetas se llaman contenedores (contienen información). Los contenedores tienen una característica de **seguridad individual** (por carpetas).

blob es un acrónimo de *objeto binario grande*.

Blobs en bloque	Blobs en páginas	Blobs en anexos
Tienen un tamaño máximo de 4,7 TB La mejor opción para <u>almacenar objetos binarios grandes y estáticos que cambian con poca frecuencia</u> Cada bloque individual puede almacenar hasta 100 MB de datos Un blob en bloques puede contener hasta 50.000 bloques	Está optimizado para admitir operaciones de lectura y escritura aleatorias. Se usa para implementar el almacenamiento en disco virtual para máquinas virtuales <u>Para blobs que requieran acceso aleatorio de lectura y escritura</u>	Su tamaño máximo es un poco más de 195 GB Es un blob en bloques que se usa para optimizar las operaciones de anexo Cada bloque individual puede almacenar hasta 4MB de datos <u>son adecuados para almacenar datos que crecen en fragmentos,</u>

e el

nen

se

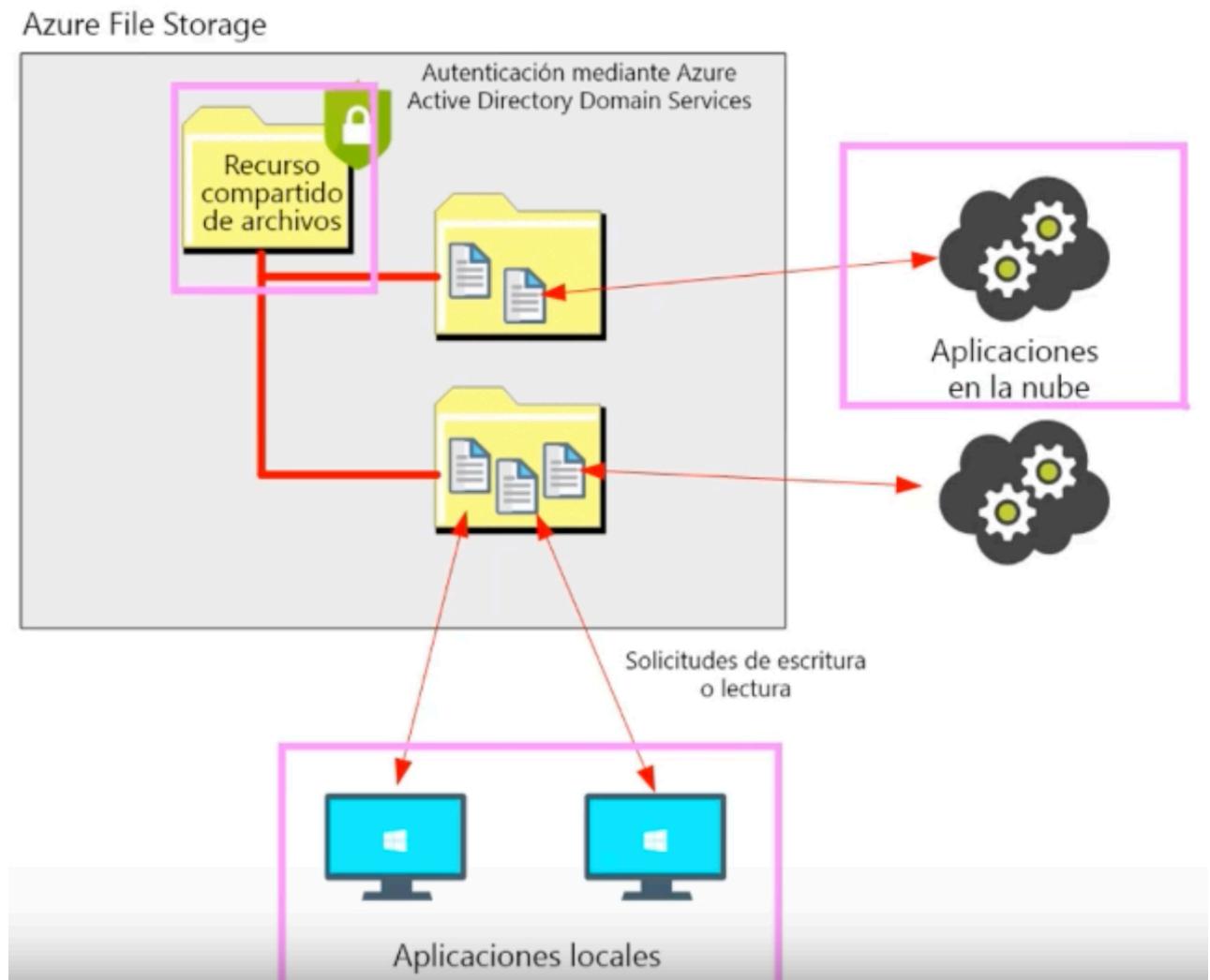
de
tos

1

- **Niveles de acceso** (vías de acceso a cada archivo/carpetas). Costo de almacenamiento vs costo de acceso
 - **Frecuente**: Acceso determinado y cuenta con alto rendimiento. Costo de almacenamiento menor vs costo de acceso menor.
 - **Esporádico**: Rendimiento inferior e incurre en gastos de almacenamiento reducidos en comparación con el nivel de acceso frecuente. Igual costo de almacenamiento vs costo de acceso.
 - **Archivo**: Proporciona el menor costo de almacenamiento pero una mayor latencia. Se almacena de forma eficaz en un estado sin conexión. Costo de almacenamiento menor vs costo de acceso.

Azure File Storage:

- Se puede guardar archivos grandes. Son recursos compartidos. Es como un SharePoint. Se le puede dar acceso a otras personas, se ven los cambios reflejados. No es una buena práctica trabajar en tiempo real con varias personas, es sensible a sobreescritura. Archivos que no cambien mucho.



acceso.

mayor vs

nparación

cencan de
so mayor.

le dar
po real

Azure Cosmos DB:

- Datos semiestructurados.
- Base de datos globalmente distribuida (sincroniza, lectura y escritura).
- Es **elástica** - de forma automática - (no manual que es escalable).
- Garantía de baja latencia (99 percentile).
- Trabaja con 5 modelos semiestructurados (tablas clave-valor → Table API, familia de columnas → Cassandra-, documentos tipo JSON → Mongo DB, grafos → Gremlin y SQL/PostgreSQL-excepción).
- ETL casi en tiempo real.

Casos de uso de Azure Cosmos DB

Web y comercio minorista

Mediante el modelo de replicación de arquitectura multamaestro de Azure Cosmos DB y los compromisos de rendimiento de Microsoft, los ingenieros de datos pueden implementar una arquitectura de datos que sea compatible con aplicaciones web y móviles que logren un tiempo de respuesta inferior a 10 ms en cualquier parte del mundo.

Juegos

El nivel de base de datos es un componente crucial de las aplicaciones de juegos. Los juegos modernos realizan un procesamiento gráfico en las consolas/dispositivos móviles los clientes, pero dependen de la nube para ofrecer contenido personalizado como estadísticas del juego, integración con redes sociales y tablas de clasificación con puntuaciones.

Escenarios de IoT

Se han diseñado y vendido cientos de miles de dispositivos conocidos como dispositivos de Internet de las cosas (IoT) para generar datos de sensores. Mediante tecnologías como Azure IoT Hub, los ingenieros de datos pueden diseñar fácilmente una arquitectura de la solución de datos que capture datos en tiempo real. Cosmos DB puede aceptar y almacenar esta información muy rápidamente.

- Escalabilidad
- Rendimiento
- Modelo de programación
- Disponibilidad

Lección 1: Prueba de conocimientos (continúa en la siguiente diapositiva)



¿Cuáles son los elementos de una clave de Azure Table Storage?

- Nombre de la tabla y nombre de la columna
- Clave de partición y clave de fila
- Número de fila

entos
ocidos
t de
atos
ogías

e
de
e



¿Cuándo debería usar un blob en bloques y cuándo debería usar un blob en páginas?

- Use un blob en bloques para datos no estructurados que requieran un acceso aleatorio para la lectura y escritura. Use un blob en páginas para los objetos discretos que apenas cambian
- Use un blob en bloques para los datos activos almacenados con el nivel de acceso frecuente, y recurra a un blob en página para los datos almacenados con un nivel de acceso esporádico o de archivo.
- Use un bloque de página para blobs que requieran acceso aleatorio de lectura y escritura. Use un blob en bloques para objetos discretos que cambian con poca frecuencia.



¿Para qué debería usar Azure File Storage?

- Para compartir archivos que se almacenen localmente con usuarios en otras ubicaciones
- Para permitir que los usuarios de diferentes ubicaciones comparten archivos
- Para almacenar archivos de datos binarios grandes que contienen imágenes u otros datos no estructurados.

Lección 1: Prueba de conocimientos (continuación)



Está creando un sistema que supervisa la temperatura en un conjunto de bloques de oficinas y configura el aire acondicionado en cada sala de cada bloque para mantener una temperatura ambiente agradable. Su sistema tiene que administrar el aire acondicionado en varios miles de edificios repartidos por el país o la región, y cada edificio suele contener al menos 100 habitaciones con aire acondicionado. ¿Qué tipo de almacén de datos NoSQL es el más apropiado para capturar los datos de temperatura y así permitir que se procesen rápidamente?

- Enviar los datos a una base de datos de Azure Cosmos DB y usar Azure Functions para procesar los datos.
- Almacenar los datos en un archivo almacenado en un recurso compartido creado con Azure File Storage.
- Escribir las temperaturas en un blob de Azure Blob Storage

8/marzo/2023:

Prácticas de Storage.

13/marzo/2023:

Exploración del análisis de datos en Azure:

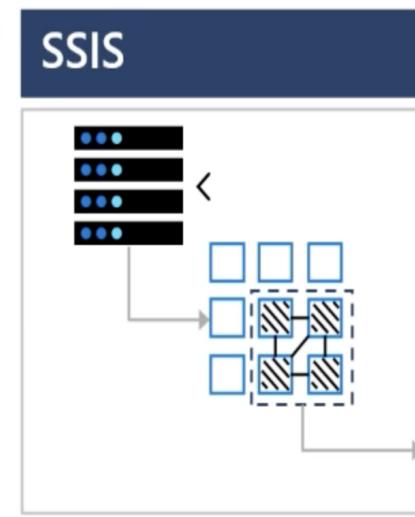
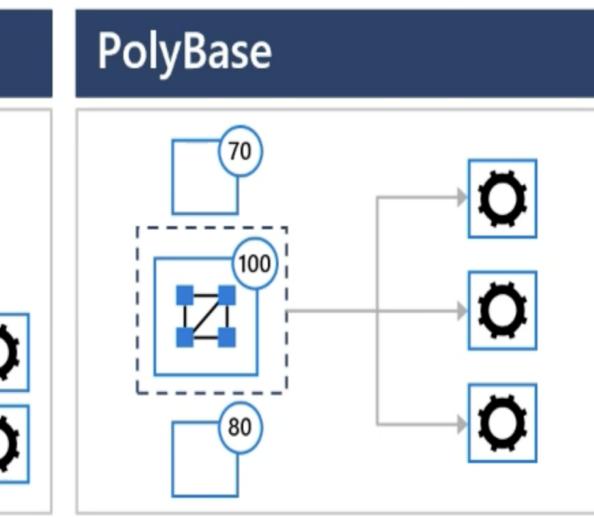
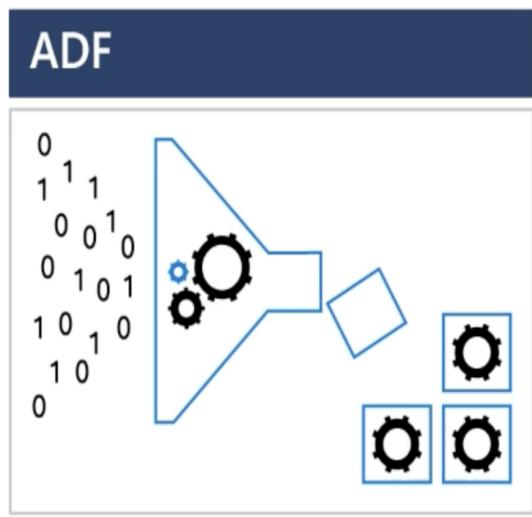
- **Azure Data Factory:** Ingestar y orquestar (paso a paso del proceso). Servicio de ingesta de datos, integración de diferentes puntos de creación de datos para llevarlos a un punto de almacenamiento escala. Es como una vía que canaliza (pipeline). Se puede hacer ETL o ELT.
- **Azure Data Bricks:** servicio de analítica (preparación, transformación y análisis). Permite procesos streaming, modelos de aprendizaje automático e Inteligencia artificial (apache Spark - librería). Es colaborativo (notebooks) con lenguajes como Python, Scala, Java y SQL. Se puede trabajar con todo

--
as
--
e
/
s
to, a
de
das las
.

librerías nativas de analítica y procesamiento. Trabaja por cluster (capacidades divididas en una misma máquina: divide, ejecuta y une – computacionalmente es más rápido).

- **Azure Synapse Analytics:** servicio integrador. Reúne la integración, almacenamiento y análisis de macrodatos. Tiene su propio servicio de ingesta. También tiene su sistema de almacenamiento y la transformación en el módulo de SQL/Spark.
- **Azure Data Lake Storage:** Repositorio para grandes cantidades de datos sin procesar. Organiza los datos en repositorios para mejorar el acceso a los archivos. Admite permisos POSIX y RBAC. Es compatible con el sistema de archivos distribuido Hadoop.

Describir la ingesta de datos en Azure:



ADF: Homogéneo

PolyBase: Basado en archivos

SSIS: Heterogéneo

15/marzo/2023:

nisma

a

os datos
ole con el

