Apuntes BD2 - Semana 6 - Viernes 2 de Septiembre

Estudiante: Andrea María Li Hernández - 2021028783

Tema principal: Data Tiers y la estimación de una base de datos plana.

Ejemplo visto en clase:

Data Tier	Number of Documents (million documents)	Expected growth 6 months (%)	Replicas	Memory Ratio RAM vs Disk	Average Document Size (KB)	Total Memory
Content	10	1	2	30	25	26.2260437
Hot	100	0	2	30	25	262.260437
Warm	600	0	2	60	25	786.781311
Cold	1200	0	0	120	25	262.260437
Frozen	3600	100	0	N/A	25	N/A

- Uno define capas de servicio que va a dar como hot, warm, cold o frozen.
- **Contenido:**
  - Son datos que no cambian frecuentemente. Por ejemplo: El ícono de una página web.
  - No se necesita poder en CPU si se trata con Contenido.
  - Ocupa poca memoria y es bastante eficiente.
- El data tier va a ser definido basado en lo que requiera el cliente.
- La principal diferencia entre los data tiers es la cantidad de archivos que va a guardar.
- En **Warm** el procesamiento no debe ser tan poderoso, la memoria sí.
- En Hot, Warm, Cold y Content: **Podemos modificar datos**.
- En Frozen **NO** podemos cambiar los datos.
  - No hay que estirar su crecimiento pues es prácticamente infinito.
  - No se usa nada de memoria.
- Cold y Frozen no ocupan réplicas porque viven en **storage barato**.
  - Es lento y confiable porque los backups están en disco.

Definir características de la BD

1. Tomar la cantidad de documentos que se van a almacenar.
2. Calcular el tamaño promedio (*Average Size*) de los documentos.
3. Calcular el *raw data* size.
4. Definir un porcentaje de espacio extra (*Extra Space*).
   - Siempre nos tenemos que dar un "colchón" para imprevistos.
5. Definir el número de réplicas:
   - Esto es bueno cuando tenemos muchas **lecturas**. Ejm:
     - 3 réplicas -> Dividir el trabajo entre 3 -> **Más performance**.

    - El problema es con las escrituras.
        - Más réplicas -> Más *overhead* de escritura.
    - El peligro de no tener réplicas es no poder levantar la base de datos al caerse, por esto se debe tener **al menos 1** réplica.
6. *Total Data* (GB): Espacio total de disco para almacenamiento.
    - Esto se define en función a la cantidad de documentos, el *average size* y la cantidad de réplicas.
7. *Total Memory*: Se define en función al almacenamiento.

El crecimiento de la base de datos se puede establecer o entender en las reglas de negocio.

- Es importante tener en cuenta que normalmente no es real o mantenible cumplirle al cliente si este desea que quiere la BD "siempre disponible".

Replicas: Cuanta información tengo de respaldo.

- Las réplicas siempre son importantes.
- Datos no tan frecuentes y tener muchas réplicas -> Quitar recursos al sistema.
- **No tiene sentido tener 1 servidor y muchas replicas:**
    - El hardware es compartido entre las réplicas, entonces no se podrían paralelizar las operaciones.
    - En el caso de que se "muera" la computadora, se perderían todas las réplicas.

Timestamp: Fecha en que entró algo al sistema o se modificó un documento

---

En **elasticsearch** los datos nunca se guardan en json, pues esto generaría *overhead*.

- **Overhead**: Todas aquellas cosas que necesito para que la computadora funcione.
    - Si el overhead es muy alto, estaríamos desperdiciando recursos.

S3 - Object Storage

Configuración de un Amazon Simple Storage Service (S3) en AWS Pricing Calculator

a) Description: Frozen Tier.

b) Elegir Region: A veces los datos ocupan meterse en una región en específico.

- GovCloud: Servicio de gobierno, el tipo de manejo de datos es diferente (diferentes regulaciones de datos).
- La región más vieja y barata es US East (N.Virginia).

c) Tipos de Storage Classes: Esto define cuánto nos cobrarán y se van a elegir dependiendo del caso de uso.

d) Definir cuántos teras de storage.

e) Volumenes.

f) Tera X Disco.

g) **Snapshots**: Es peligroso hacer snapshots a nivel de disco en una base de datos, pues la transacción puede quedar a incompleta y crear inconsistencias. Uno normalmente desactiva los snapshots, porque uno asumirá la responsabilidad.

Instance Types

Todos los cloud providers tienen Instance Types.

- Instance size predefinido.
- Cantidad de CPU virtuales: Threads a nivel de CPU.
- Cantidad de memoria.
- Tipo de storage que puedo utilizar.
- Network bandwidth: Máxima cantidad de datos por unidad de tiempo.

Tipos de Instancias

1. **General Purpose**: Balance de recursos.
2. **Compute Optimized**: Muchos *threads* y memoria baja.
3. **Memory Optimized**: Mucha memoria con respecto a la cantidad de procesadores que vamos a tener.
4. **Storage Optimized**: Acceso a disco rápido.
5. **Accelerated Computing**: Muchas GPUs y se suele usar en machine learning.

EC2

- **Master Nodes:**
  - Máquinas simples.
  - Hacen tareas.
  - El disco no es relevante.
- Si entiendo el acceso a los datos, puedo definir el hardware que va a necesitar la BD.

Otros conceptos

- Ocupamos **telemetría/observabilidad** para ir viendo la configuración de la base de datos.
- **Utilization**: Estimar cuánto tiempo estará corriendo la máquina.
- Usualmente la documentación de la bases de datos no recomienda tener **shards** de más de 100 GB.
- **Load test**: Para comprobar el comportamiento apropiado de la BD.
- Con las bases NoSQL, varias instancias pueden vivir dentro de la misma máquina; pero todos van a competir por el mismo hardware.

Alta prioridad

- **Manage Services**: Una empresa vende las BD ya instaladas y se encargan del mantenimiento de estas.
- **SAAS**: Software As A Service.
- **PAAS**: Platform As A Service.