> *Bases de Datos 2, Apuntes*

https://api.biorxiv.org/covid19/0 -> data set para hacer requests para jalar la información

Cursor -> desplazamiento en el archivo

Count -> cantidad que jala

Total -> cantidad de datos

Schema -> nombre de las columnas que son los atributos en el JSON

- Son datos semi estructurados.
- Se pueden pasar a tablas SQL, pero pueden perder consistencia.

Para optimizar el storage en NoSQL, se puede repetir información para búsquedas más rápidas. La información está en un solo documento en lugar de muchas tablas.

Esto puede pasar cuando se tiene un motor NoSQL y dos entidades que luchan por un recurso compartido y ambos piensan que tienen la exclusividad del recurso. Una solución es usar una tabla con los documentos y con una columna de estado para verificar el recurso. Esto funciona, pero va en contra de las reglas de normalización. Tambíen, puede pasar en data warehouse, en el cual la información se encuentra en una sola tabla.

Semántica de los datos -> significado que tienen las palabras en el contexto.

Las palabras Gerardo, Jerardo, Gerard0 y Gerard son distintos, pero todos tienen el mismo significado.

El IA-NLP identifica si los datos representan lo mismo.

La base de datos no puede correr el IA-NLP, por lo que se ocupa de una lógica que garantice que los datos esten preprocesados antes de almacenarlos a la base de datos.

Estandares de la base de datos -> una forma de autodocumentar la base de datos.

Ejemplo : fk_nombreTabla_campoTablaFuente_campoTablaDestino -> fk_inst_idInst_idInst

DOI -> número único de un artículo científico de tipo string con 128 caracteres.

Long text -> texto de cualquier tamaño.

Para hacer una búsqueda de texto completo, la base de datos hace una estructura como un índice invertido.

Link -> string de 256 caracteres porque el protocolo lo define. Está relacionado con el DOI y publish site, por lo que no se pone en la mismo tabla porque tiene dependencia funcional entre ellos. Esto puede ocasionar que se muestren datos antiguos. Una solución es crear un campo template con los caracteres repetidos del link.

Crear una base de datos en nube

VPC -> virtual private cloud

- crea un servicio en red

- tags -> labels (clave valor)

Se crean subredes asociados a un VPC:

- Públicos: DMZ -> zona desmilitarizada. Se expone a internet la zona a ser hackeada.
- Privados: recursos más importantes como datos. Es restringido, por lo que no se permite el acceso a cualquiera. El acceso de público a privado tiene restricción de IP y RBAC (Role Based Access Controls)

Permite regionalizar la red -> availability zone, como data centers.

Mínimo de 3 data centers por la forma en que se elige el nodo master.