

Almacenamiento y captura de datos

Claudio Aracena

GobLab - Universidad Adolfo Ibáñez
Chatbot Chile



Contenidos

- Captura de datos desde archivos
- **Base de datos**
- Captura y almacenamiento de datos en BD
- Captura de datos de la Web (Web scraping)
- Captura de datos de API (ej: Twitter)
- Captura y almacenamiento en arquitecturas Big data

Códigos y clase en:

<https://github.com/caracena/almacenamiento-captura-datos>

Clase de hoy



Base de datos

- Revisión consultas clase anterior
- Base de datos en servidor
- Base de datos no relacionales





Google Cloud Platform

Cloud SQL

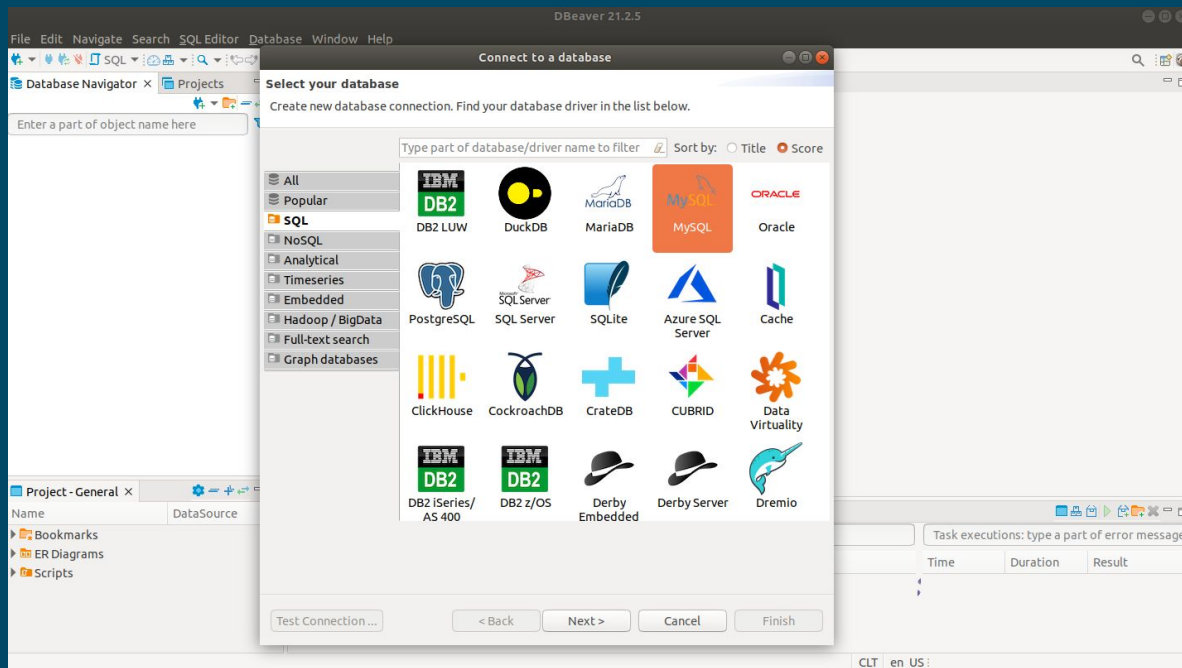
- Base de datos alojada en la nube de Google <https://cloud.google.com/>
- 3 opciones de motores (SGBD) de base de datos
 - MySQL
 - PostgreSQL
 - SQL Server
- Se paga por cpu, ram, espacio, tipo de espacio, ubicación de los servidores, entre otros

Dbeaver



- Cliente SQL

- <https://dbeaver.io/>





Base de datos no relacionales (NoSQL)

NoSQL es el término genérico usado para referirse a almacenamiento de datos que no sigue el modelo tradicional de base de datos relacionales. Específicamente, la data no sigue el modelo entidad-relación y no utiliza SQL como lenguaje de consulta.

Ejemplos de estas bases de datos son

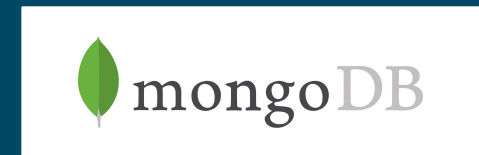
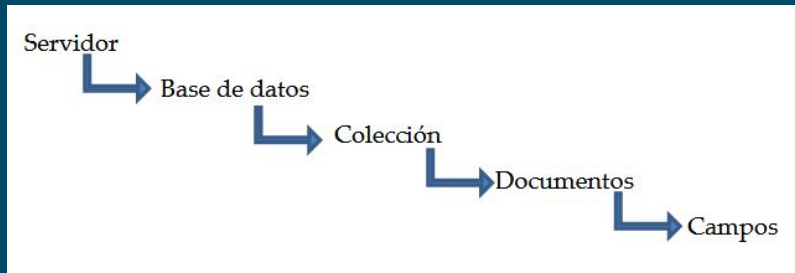
- MongoDB (document-oriented)
- Cassandra, Hbase (column-oriented)
- Redis (key-value)
- Neo4j (graph-oriented)





Base de datos documentales

- Una base de datos documental está constituida por un conjunto de programas que almacenan, recuperan y gestionan datos de documentos o datos de algún modo estructurados.
- A diferencia de las bases de datos relacionales, estas bases de datos están diseñadas alrededor de una noción abstracta de "Documento".
- En MongoDB un documento es un conjunto de datos almacenado en formato JSON



MongoDB



Comparación entre MongoDB y base relacional

TFN	Name	Email	age
12345	Joe Smith	joe@gmail.com	30
54321	Mary Sharp	mary@gmail.com	27

} two rows

```
{ _id: 12345,  
  name: "Joe Smith",  
  email: "joe@gmail.com",  
  age: 30  
}
```

```
{ _id: 54321,  
  name: "Mary Sharp",  
  email: "mary@gmail.com",  
  age: 27  
}
```

} two documents

MongoDB



Comparación entre MongoDB y base relacional

```
{_id: 12345,  
  name: "Joe Smith",  
  emails: ["joe@gmail.com", "joe@ibm.com"],  
  age: 30  
}  
  
{_id: 54321,  
  name: "Mary Sharp",  
  email: "mary@gmail.com",  
  age: 27  
}
```

TFN	Name	Email	age
12345	Joe Smith	joe@gmail.com , joe@ibm.com ??	30
54321	Mary Sharp	mary@gmail.com	27

MongoDB



Comparación entre MongoDB y base relacional

```
{ _id: 12345,  
  name: "Joe Smith",  
  email: ["joe@gmail.com", "joe@ibm.com"],  
  age: 30  
}  
  
{ _id: 54321,  
  name: "Mary Sharp",  
  email: "mary@gmail.com",  
  age: 27,  
  address: { number: 1,  
             name: "cleveland street",  
             suburb: "chippendale",  
             zip: 2008  
            }  
}
```

TFN	Name	Email	age	address
12345	Joe Smith	joe@gmail.com	30	
54321	Mary Sharp	mary@gmail.com	27	1 cleveland street, chippendale, NSW 2008



MongoDB

Comparación entre MongoDB y base relacional

ID	Nombre	Email	Edad
11111	Juan	juan@udd.cl	30
2222	Carlos	carlos@udd.cl	35

```
{_id: "11111", nombre: "Juan", email: "juan@udd.cl", edad: 30}  
{_id: "2222", nombre: "Carlos", email: "carlos@udd.cl", edad: 35}
```

<https://www.jdoodle.com/online-mongodb-terminal/>

<https://www.humongous.io/app/playground/mongodb/new>



Base de datos basadas en grafos

- Representa la información como nodos de un grafo y sus relaciones con las aristas del mismo
- **Ventajas**
 - Consultas realmente rápidas cuando busca relaciones entre nodos
 - Realmente rápido para recorrer nodos
 - Puede representar múltiples dimensiones
- **Desventajas**
 - Inapropiado para información transaccional, como registros contables donde las relaciones entre registros son más simples
 - Es difícil hacer consultas agregadas de manera eficiente

<https://sandbox.neo4j.com/>





Base de datos llave-valor

- Guardan tuplas que contienen una clave y su valor.
- Cuando se quiere recuperar un dato, simplemente se busca por su clave y se recupera el valor.
- En general es para tipos de datos simples, o cuando queremos buscar un dato en particular en forma rápida
- **Desventajas:**
 - No es muy útil para almacenar relaciones.
 - Es difícil mantener llaves únicas cuando los datos aumentan

<https://try.redis.io/>





Base de datos columnares

- Una base de datos en columnas está optimizada para lograr una recuperación rápida de columnas de datos
- Normalmente son usadas en aplicaciones analíticas.
 - Para realizar operaciones de agregación (min, max, mean) sobre datos en particular
 - Esto es eficiente pues recupera toda la columna en forma rápida



Clase de hoy



Base de datos

- Revisión consultas clase anterior
- Base de datos en servidor
- Base de datos no relacionales

