Bases de Datos II
Sesión 0 — Introducción

Fernando Terroso Sáenz

Departamento de Tecnologías de la Información y las Comunicaciones
Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Politécnica de Cartagena

fernando.terroso@upct.es

2024

CURSO ACADÉMICO	2023/2024
TITULACIÓN	GRADO EN CIENCIA E INGENIERÍA DE DATOS
CUATRIMESTRE	SEGUNDO
CURSO	SEGUNDO
CARÁCTER	OBLIGATORIA
CRÉDITOS ECTS	6
DEPARTAMENTO	INGENIERÍA Y TECNOLOGÍA DE COMPUTADORES

Profesor y horario

	Profesor	Horario	Aula
Teoría	Fernando Terroso	Miércoles 9:00–11:00	1.6
Prácticas	Isaac Martínez	Miércoles 18:00–20:00	Inf-3

Tutorías	Lunes 11:00–13:00 (y tutorías electrónicas)
Contacto	Despacho 21 fernando.terroso@upct.es

Introducción a la asignatura

Tópicos

- Modelado de datos NoSQL y NewSQL
- Lenguajes de consulta de datos
- Uso de bases de datos Key/Value, Documentales, Columnares y de Grafos
- Formatos de archivos de intercambio de datos (CSV, XML, JSON, Avro)
- Visualización de datos (introductorio)
- Dimensionamiento correcto de las soluciones de tratamiento de datos

Data Access Hitting a Wall

Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.

- Oh!, and 1PB ~5,000 disks

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where databases can help

- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1$)
- … 2 days and 1K$
- … 3 years and 1M$

[slide src: Jim Gray]

Data Science

- Se pasa un **80–90** % del tiempo importando datos, organizando para la optimización, etc.
- Estudiaremos las **abstracciones** del **modelado de datos**, así como las herramientas (**bases de datos**) disponibles
  - **¿cómo se organizan y modelan los datos?**
  - **¿cómo se consultan de forma eficiente?**
  - **¿cómo se tratan los datos heterogéneos?**
  - **¿cómo se gestionan cantidades de datos que no caben en una máquina (*big data*)?**
  - **Modelado relacional, índices, desnormalización, duplicación, agregación**
  - …
- Estudiaremos modelado de datos NoSQL y NewSQL

"…no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics."

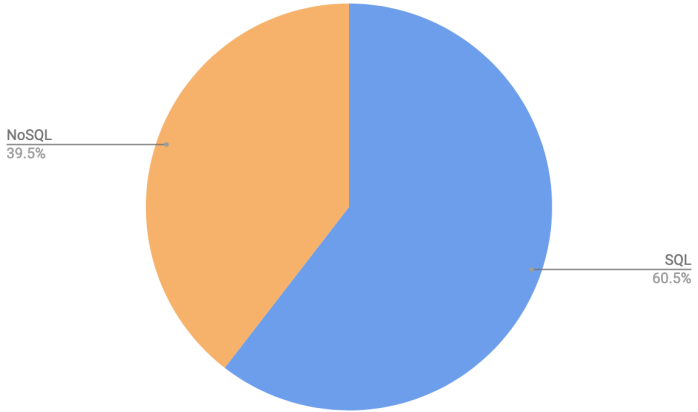
— Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", Gartner, 2001

¿Bases de datos hoy?

- No hay fuentes fiables que muestren el uso real de las bases de datos
- Hay sitios web como *Database Ranking*, `https://db-engines.com/en/ranking`
- Otros: Encuesta en DeveloperWeek 2019 (SF/Bay), 8000+ desarrolladores[1]
- También: Stackoverflow Devloper Survey 2019[2]

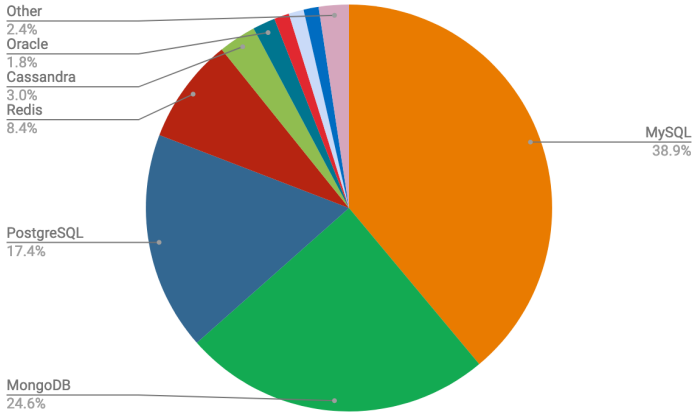
¿Bases de datos hoy? (cont.)

- SQL vs. NoSQL:



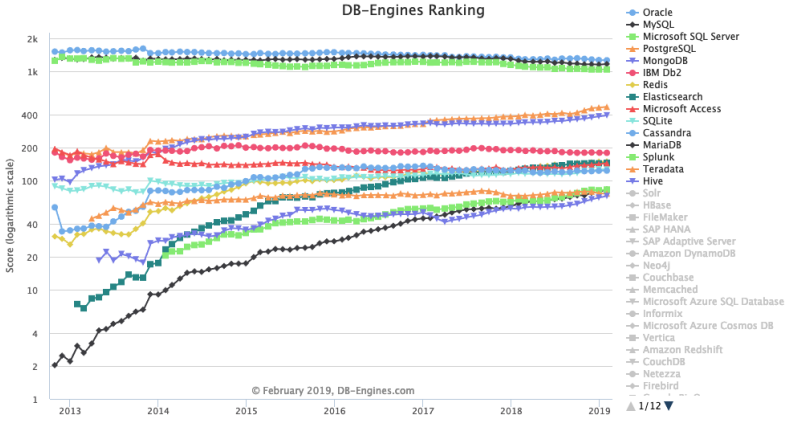
¿Bases de datos hoy? (cont.)

- BBDDs más populares:



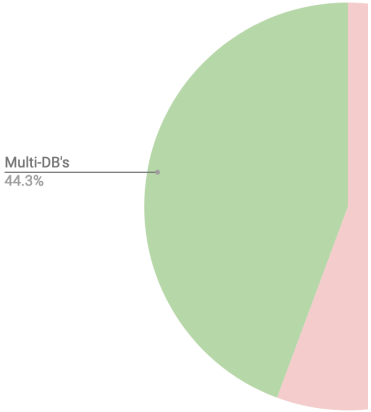
¿Bases de datos hoy? (cont.)

- Comparado con DB-Ranking (`https://db-engines.com/en/ranking`):



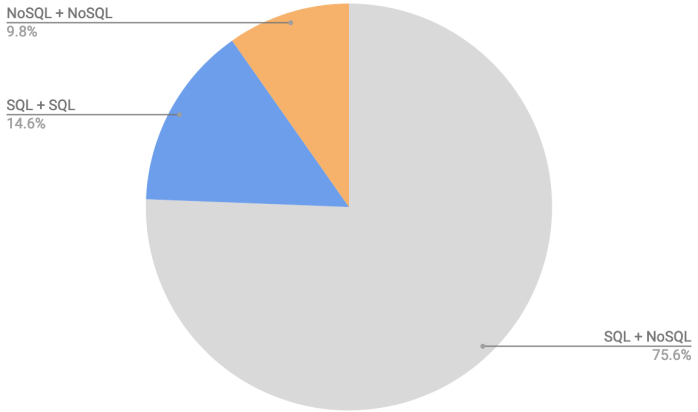
¿Bases de datos hoy? (cont.)

- Multi-Database (persistencia políglota):



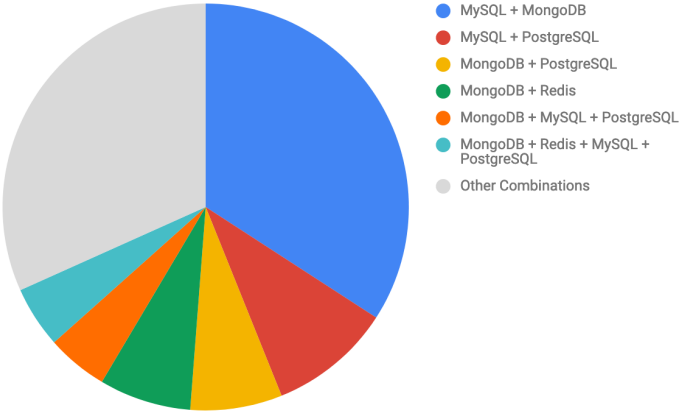
¿Bases de datos hoy? (cont.)

- Combinaciones SQL/NoSQL:



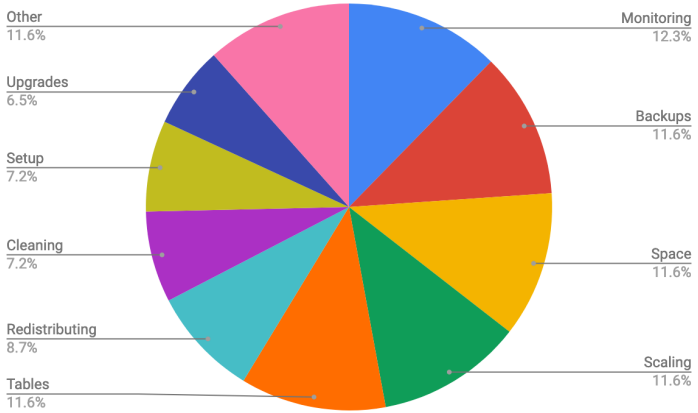
¿Bases de datos hoy? (cont.)

- Combinaciones más populares:



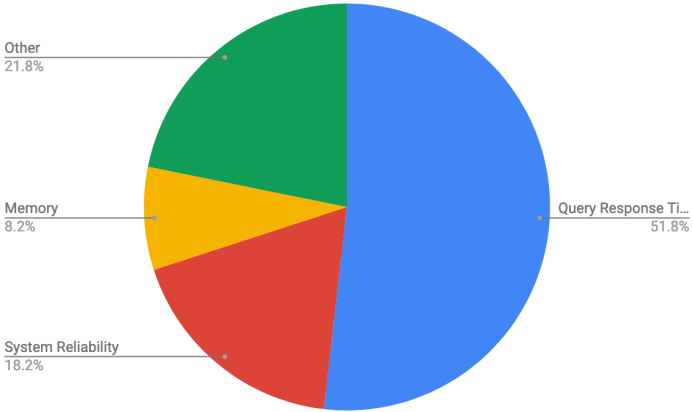
¿Bases de datos hoy? (cont.)

- Tareas que más consumen tiempo:



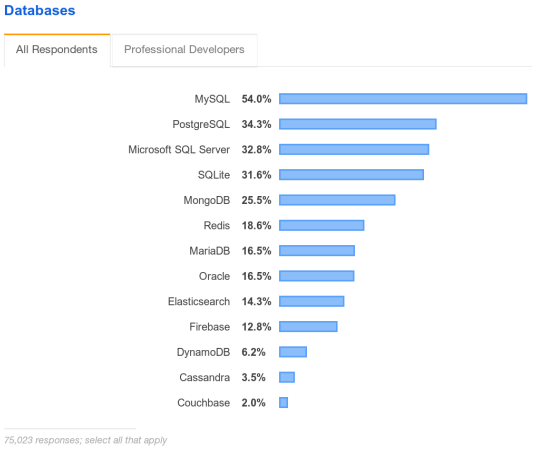
¿Bases de datos hoy? (cont.)

- Métricas más importantes hechas en la BD:



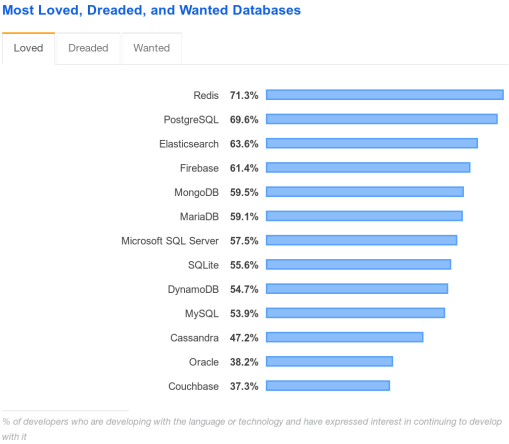
¿Bases de datos hoy? (cont.)

- Stackoverflow: Uso de bases de datos



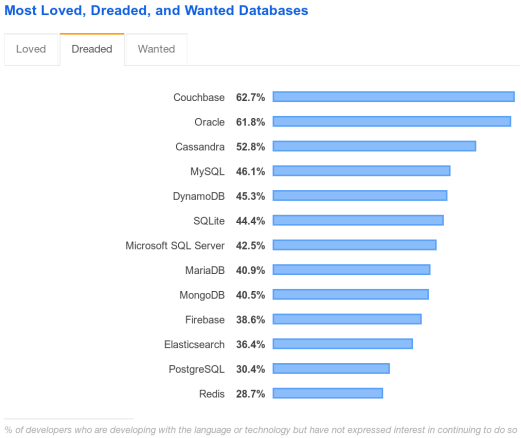
¿Bases de datos hoy? (cont.)

- Stackoverflow: Bases de datos más queridas



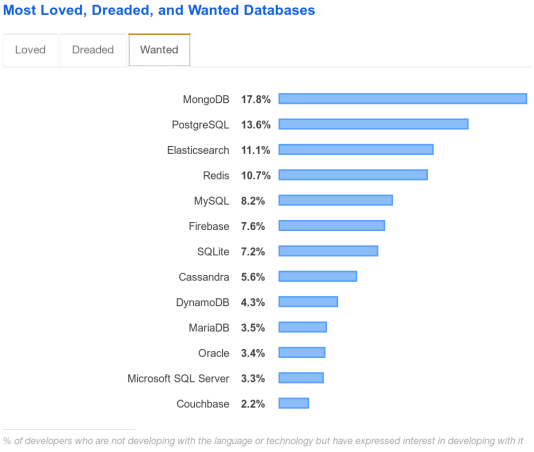
¿Bases de datos hoy? (cont.)

- Stackoverflow: Bases de datos más odiadas



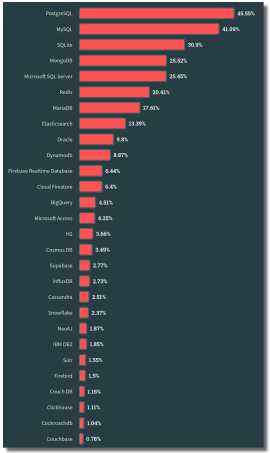
¿Bases de datos hoy? (cont.)

- Stackoverflow: Bases de datos más deseadas



¿Bases de datos hoy? (cont.)

- Stackoverflow: Bases de datos, encuesta de 2023



¿Bases de datos hoy? (cont.)

- Fuentes:

[1] http://highscalability.com/blog/2019/3/6/2019-database-trends-sql-vs-nosql-top-databases-single-vs-mu.html, tomando como fuente https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/.

[2] https://insights.stackoverflow.com/survey/2019.

Planificación del curso

Teoría	Prácticas	Fecha
Introducción a la asignatura	–	*Sem. 22/01*
Recuperación de datos y formatos de serialización	Intro & Pr. de recuperación	*Sem. 29/01*
Introducción a NoSQL	Pr. de recuperación (ii) (*)	*Sem. 05/02*
Introducción a NoSQL (ii)	–	*Sem. 12/02*
Documentos (diseño vs. sql)	Práctica documentos	*Sem. 19/02*
Documentos (mongodb intro y consultas)	Práctica documentos (ii)	*Sem. 26/02*
Documentos (mongodb agregación, índices, etc.)	Práctica documentos (iii) (*)	*Sem. 04/03*
Clave-valor	Práctica clave valor (*)	*Sem. 11/03*
NoSQL: Bases de datos columnares	NoSQL: Bases de datos columnares	*Sem. 18/03*
Repaso y ejercicios	Repaso prácticas	*Sem. 01/04*
NoSQL: Bases de datos columnares (ii)	NoSQL: Bases de datos columnares (ii) (*)	*Sem. 08/04*
NoSQL: Bases de datos de grafos	Práctica BBDD de Grafos	*Sem. 15/04*
NoSQL: Bases de datos de grafos	Práctica BBDD de Grafos (*)	*Sem. 22/04*
Repaso y ejercicios	Finalización de prácticas	*Sem. 29/04*

(*) Entrega de prácticas

Evaluación

- Prácticas de la asignatura: 60 %
  - 30 % Entregas semanales
    - Notebooks realizados en las sesiones de prácticas
    - Se abrirá tarea para entregarla **el mismo día**
  - 30 % Parte de prácticas en examen final
- Examen final teórico-práctico: 40 %

Bibliografía

- Nathan Marz, James Warren. **Big Data: Principles and best practices of scalable realtime data systems**, Manning Publications, 2015
- Eric Redmond, Jim R. Wilson. **Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement**. Pragmatic Bookshelf, 2012
- Pramod J. Sadalage, Martin Fowler. **NoSQL Distilled. A Brief Guide to the Emerging World of Polyglot Persistence**. Addison-Wesley, 2013

Repositorio de la asignatura

- Guiones de prácticas e información adicional
- El repositorio está alojado en GitHub y se llama ‘`bd2-public`’, dirección `https://github.com/dsevilla/bd2-public`
- Para obtenerlo (rama **23-24**):

```
$ git clone https://github.com/dsevilla/bd2-public.git
$ cd bd2-public
```

- No hace falta bajarlo porque usaremos **Google Colab**
- (Esto requiere una cuenta Google)
- Los *Notebooks* se podrán guardar en **Drive** o en un repositorio **GitHub** y luego enviar al profesor

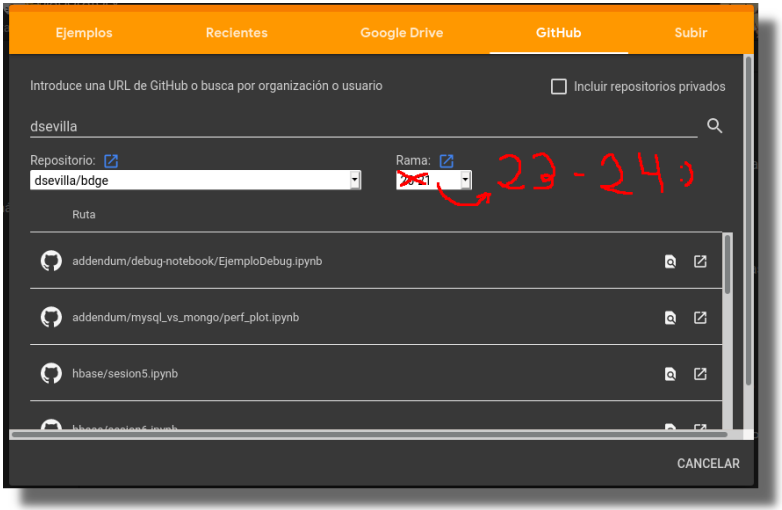
Google Colab

- https://colab.research.google.com/

Google Colab (cont.)

- En `Archivo→Abrir Cuaderno`
- Seleccionar **GitHub**
  - **Habrá que dar permiso de acceso, sólo repositorios públicos si no queréis dar acceso a vuestros repositorios privados**
- En el primer cuadro de texto: **dsevilla**
- Repositorio: **dsevilla/bd2-public**, rama: **23-24**
- Ruta: Sesiones alojadas en subdirectorios (p. ej. **intro/sesion0.ipynb**)
- Al terminar de completar el *notebook*, `Archivo→Guardar Cuaderno`
  - **Drive**, ó
  - **GitHub** (repositorio vuestro, puede ser privado)

Google Colab (cont.)



Google Colab (cont.)