Bases de Datos

Introducción

Dr. Gerardo Rossel

1 Cuat. 2025

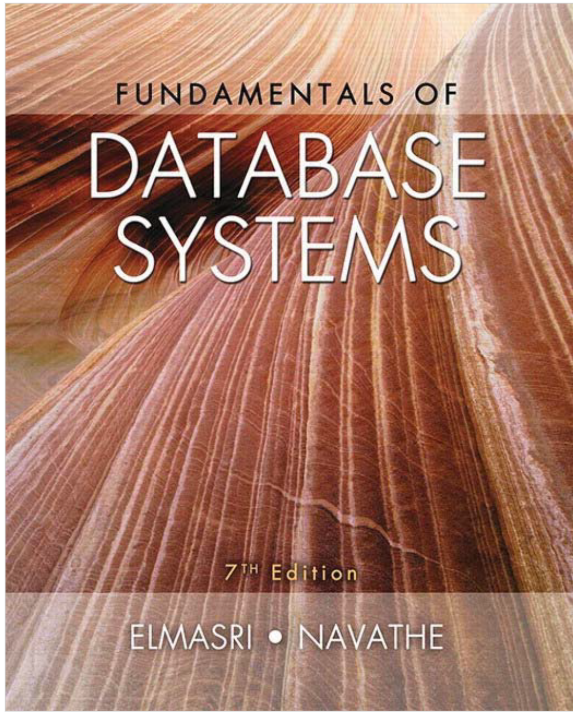
¿Qué es una base de datos?

Definición

Elmasri-Navathe

Una **base de datos** es una colección de datos relacionados.

Datos : Hechos conocidos que pueden ser registrados y que tienen significado implícito.



Propiedades clave de una Base de Datos

- **Representación del mundo real**: Refleja un aspecto específico de la realidad, conocido como **mini mundo** o **universo de discurso (UoD, Universe of Discourse)**.
- **Coherencia lógica**: Los datos almacenados tienen una estructura organizada y un significado inherente.
- **No es un conjunto aleatorio**: Una base de datos debe estar estructurada; un simple conjunto de datos sin relación no se considera una base de datos.

Atención

Una base de datos tiene una fuente de la cual se derivan los datos, algún grado de interacción con eventos en el mundo real y una audiencia que está activamente interesada en su contenido.

Sistema de Gestión de Bases de Datos(DBMS)

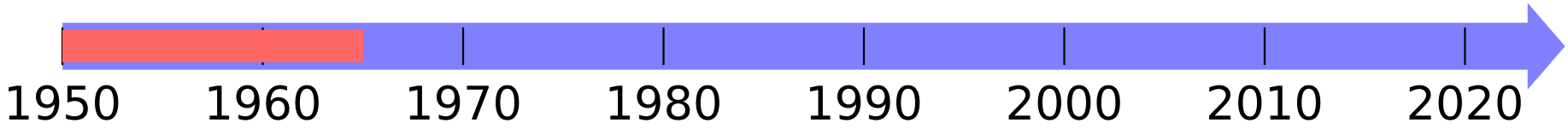
- Un DBMS es un sistema (de software) para el manejo y administración de bases de datos
- El DBMS es un sistema de software de propósito general que facilita los procesos de definición, construcción, manipulación y distribución de bases de datos entre varios usuarios y aplicaciones.
  - Capacidad de administrar datos persistentes.
  - Habilidad de acceder a grandes cantidades de datos de manera eficiente.
  - Soporte para al menos un modelo de datos, permitiendo vistas abstractas a los usuarios.
  - Soporte para lenguajes de alto nivel para definir, estructurar y manipular datos, y el acceso a datos (consultas).

DBMS

- Permite a los usuarios crear nuevas bases de datos y especificar sus esquemas.
- Brinda a los usuarios la capacidad de consultar y modificar los datos.
- Soporta el almacenamiento de cantidades muy grandes de datos durante un largo período de tiempo, permitiendo un acceso eficiente a los mismos.
- Permite la **durabilidad** , es decir, la recuperación de la base de datos ante fallos, errores de diversos tipos o uso indebido intencional.
- Controla el acceso a los datos de muchos usuarios a la vez:
    - Sin permitir interacciones inesperadas entre los usuarios (**aislamiento**)
    - Sin que las acciones sobre los datos se realicen parcialmente (**atomicidad**)

Algo de Historia

Comenzamos

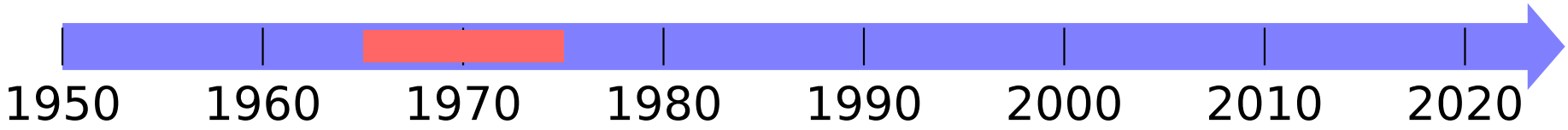


- Almacenamiento de datos en cintas magnéticas
- Para procesar datos, se leía el input de cintas y se persistía el output en una nueva cinta
- El acceso era secuencial , y los datos estaban en un único orden

Atención

Era muy importante el orden de los datos para agilizar el procesamiento

LLegaron los discos

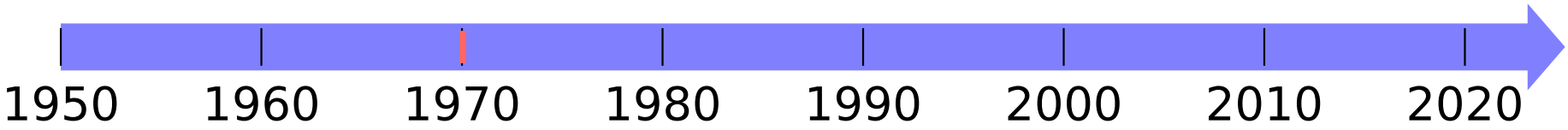


- Almacenamiento de discos rígidos, con acceso directo a los datos.
- Surgen distintos modelos de datos para describir la estructura:
  - **Jerárquicos** : basados en árboles. IBM IMS se sigue usando en la actualidad en mainframes.
  - **De red** : basados en grafos, estandarización CODASYL a fin de los 60s). CA IDMS se sigue usando en la actualidad en mainframes
- Los sistemas no soportaban lenguajes de consulta de alto nivel.

Atención

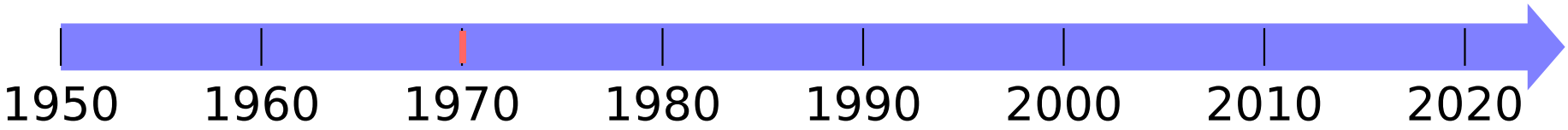
Los programadores tenían que visualizar los datos tal como estaban almacenados, y programar la manera de saltar de un dato a otro, con mucho detalle

Aparece CODD



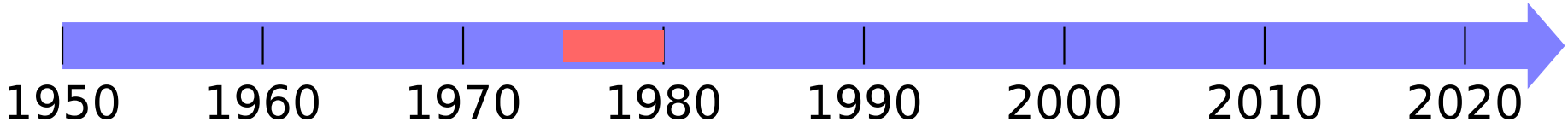
- El artículo de Edgar F. Codd titulado “**A Relational Model of Data for Large Shared Data Banks”**, publicado en 1970, es uno de los trabajos más influyentes en la historia de la informática.
- Codd introdujo el concepto de **independencia lógica** y **física** de los datos
- Este *paper* dió origen a las bases de datos relacionales

Aparece CODD



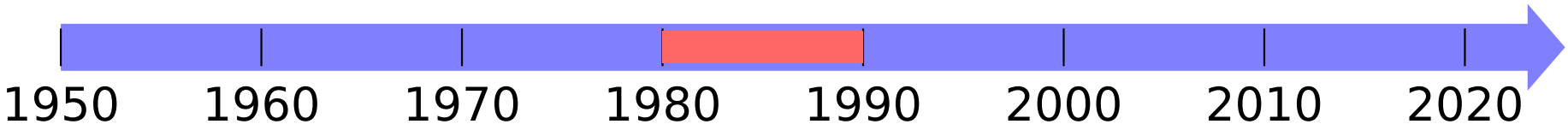
- El artículo de Edgar F. Codd titulado “**A Relational Model of Data for Large Shared Data Banks”**, publicado en 1970, es uno de los trabajos más influyentes en la historia de la informática.
- Codd introdujo el concepto de **independencia lógica** y **física** de los datos
- Este *paper* dió origen a las bases de datos relacionales

Surgen las BBDD Relacionales

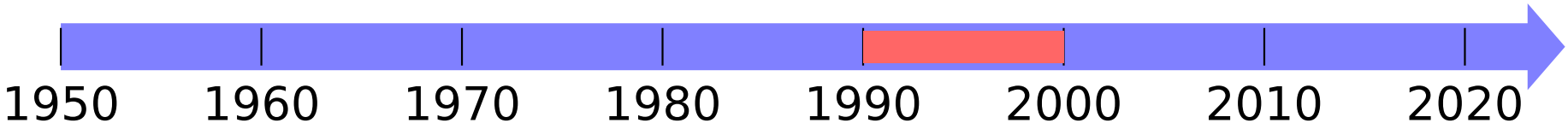


- Basados en paper de Codd, con una base teórica muy sólida.
- Los sistemas presentan a los usuarios vistas de los datos organizados como tablas, que implementan relaciones (en el sentido de la Teoría de Conjuntos)
- Las consultas de los datos se realizan con un lenguaje de alto nivel declarativo : el más importante es SQL Structured Query Language
- Los desarrolladores no tienen que preocuparse de la estructura de almacenamiento de los datos.

Dominio Absoluto de las BBDD Relacionales



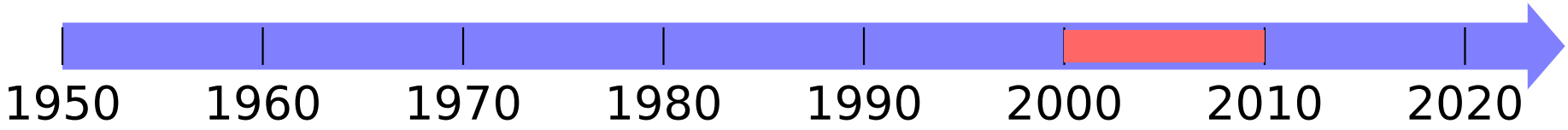
- Gran performance y facilidad de uso.
- Propiedades ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad)
- El uso principal era el procesamiento transaccional (OLTP), intensivo en actualizaciones, por lo general a nivel de transacción individual.
- 1986: primer estándar **SQL** (Structured Query Language)



- Los **RDBMS** eran el estándar de hecho para las organizaciones.
- **1992**: primera revisión importante del estándar **SQL**, adquiere mayor madurez.
- Se desarrollaron Bases de Datos orientados a objetos (OODBMS)
- Existían muchas aplicaciones, datos globales contenidos en varias BBDD, necesidad de contar con información para tomar decisiones, se requiere integración de información:
  - Surgen los Data Warehouses
  - El uso principal es el procesamiento analítico (OLAP), intensivo en consultas con grandes volúmenes

Atención

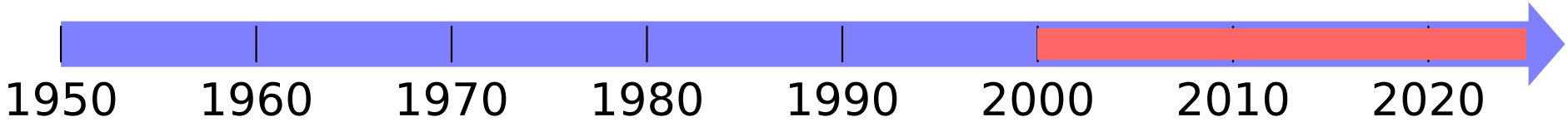
En esta década crece la World Wide Web de manera explosiva, impactando en los requerimientos sobre las bases de datos: muy altas tasas de procesamiento transaccional, disponibilidad 7 x 24, etc.



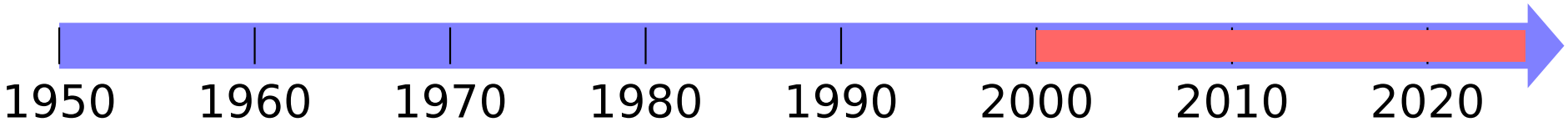
- Los RDBMS comienzan a soportar datos semiestructurados (XML y JSON) y datos espaciales.
- Teorema CAP. Eric A. Brewer. *Towards robust distributed systems*
- *MapReduce: Simplified Data Processing on Large Clusters*. Jeffrey Dean and Sanjay Ghemawat - 2004
- Google publica la arquitectura de Bigtable Chang, Dean, Ghemawat, et al. *Bigtable: A Distributed Storage System for Structured Data*. 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI) (2006)
- DeCandia, Hastorun, Jampani, et al. *Dynamo: Amazon's Highly Available Key-Value Store* 2007
- *The End of an Architectural Era (It's Time for a Complete Rewrite)*. Stonebraker, Hachem, Helland

Not Only SQL

El nombre NoSQL tiene su origen en una reunión, para los investigadores y empresas que estaban trabajando en los nuevos modelos emergentes de bases de datos no relacionales, realizada en San Francisco en el año 2009 y organizada por Johan Oskarsson.



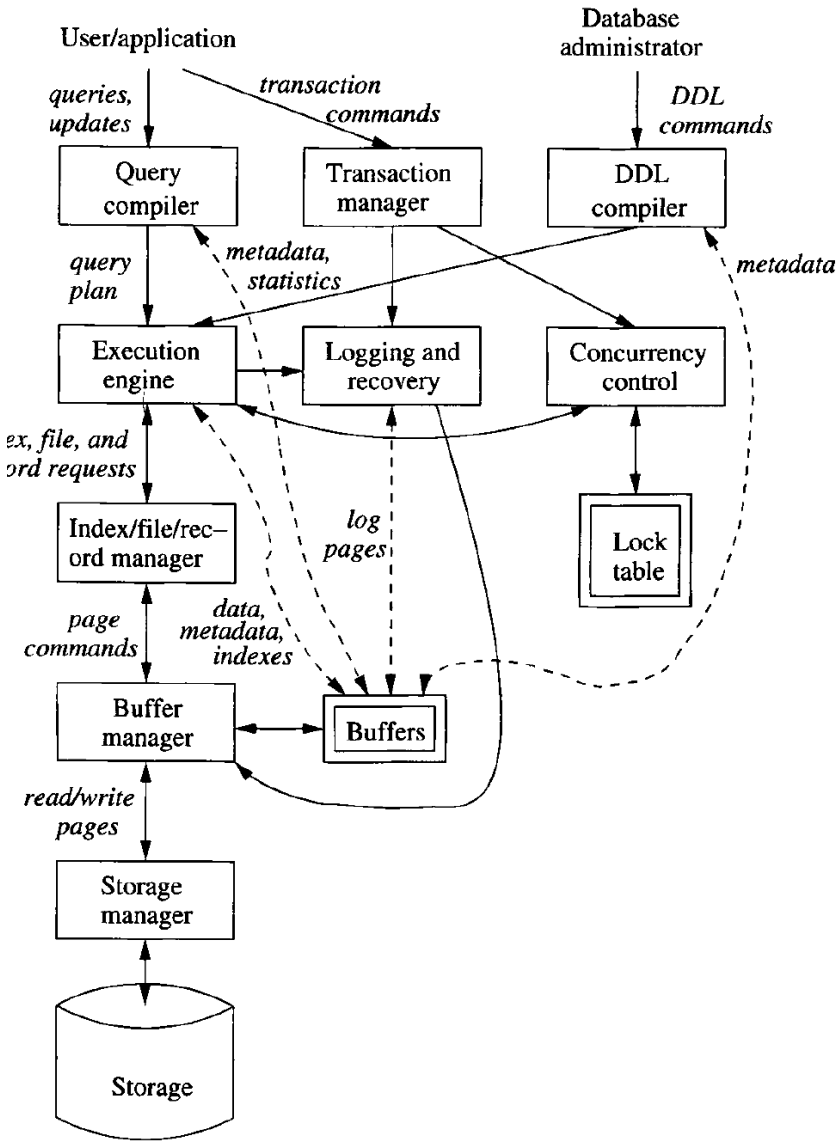
- **SQL Estándar 2023**
  - **Consultas de grafos (SQL/PGQ):** Se añadió una nueva parte al estándar que permite trabajar con datos en forma de grafos, facilitando la consulta de relaciones complejas sin necesidad de unir múltiples tablas.
  - **Tipo de dato JSON:** Se introdujo un tipo de dato nativo para JSON, mejorando la manipulación y almacenamiento de datos en este formato dentro de las bases de datos SQL.
  - **Funciones de manipulación de cadenas:** Se agregaron funciones como `LPAD()` y `RPAD()` para rellenar cadenas con caracteres específicos, facilitando tareas de formateo de texto.



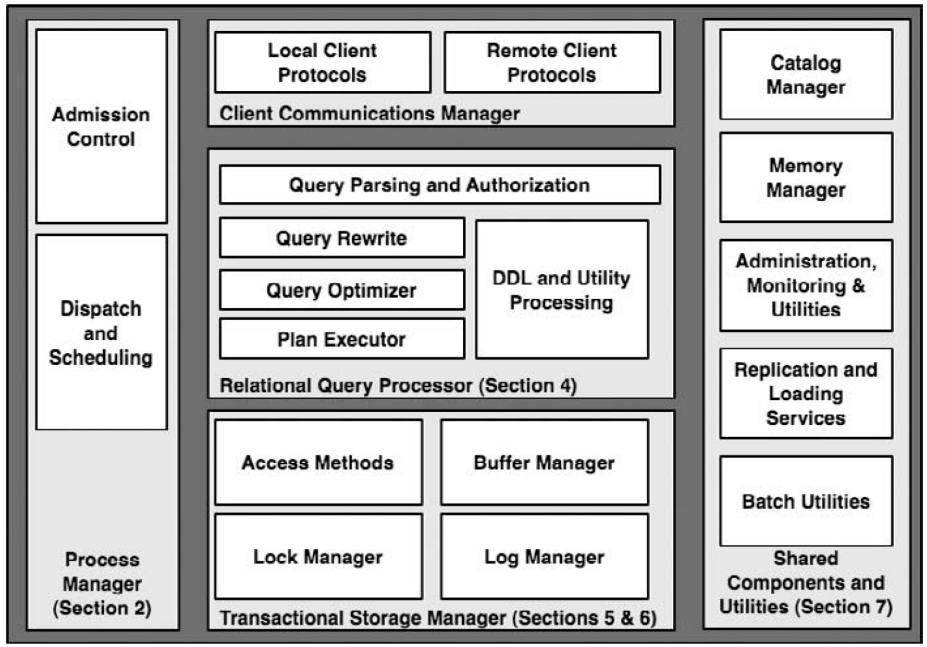
- Se profundiza el almacenamiento en la nube.
- Surgen las Bases **New SQL**: aparecen para proveer escalabilidad, performance y cumplimiento de propiedades ACID de las transacciones.
  - SQL as the primary interface.
  - ACID support for transactions
  - Non-locking concurrency control.
  - High per-node performance.
  - Parallel, shared-nothing architecture

Arquitectura de un DBMS

Arquitectura



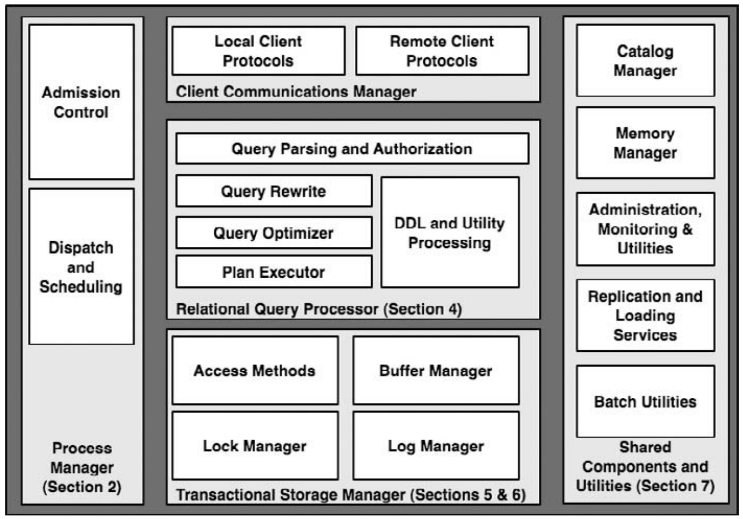
García Molina - Ullman Widom
"Database Systems: The Complete Book"



M. Hellerstein, Michael Stonebraker, and James Hamilton
" Architecture of a Database System"

Una consulta

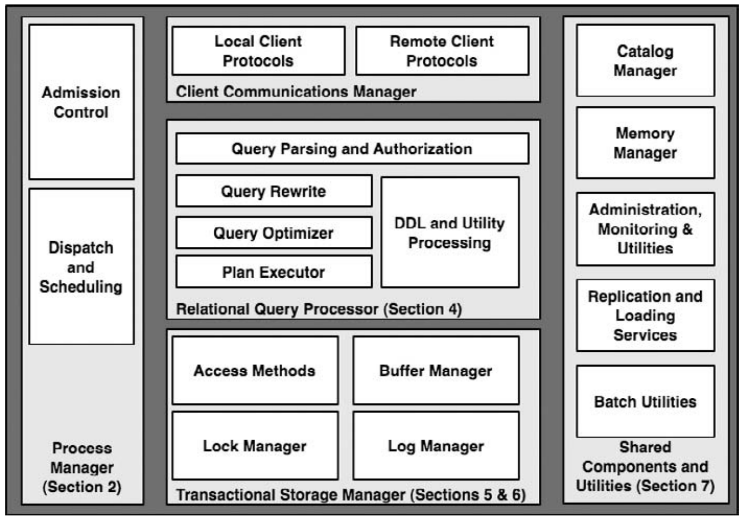
1. La aplicación llama a una API que se comunica con el **Client Communication Manager.** El CCM se encarga de establecer y recordar el estado de la conexión para el llamante (ya sea un cliente o un servidor de middleware), responder a los comandos SQL del llamante y devolver tanto datos como mensajes de control (códigos de resultado, errores, etc.) según corresponda.
2. Al recibir la consulta el **Process Manager** asigna un hilo de ejecución.También debe asegurarse de que las salidas de datos y control del hilo estén conectadas a través del gestor de comunicaciones con el cliente. Realiza además el **control de admisión**



Joseph M. Hellerstein, Michael Stonebraker, and James Hamilton
“ Architecture of a Database System”

Una consulta

3. Admitida la consulta se llama al **Relational Query Processor**. Este conjunto de módulos verifica que el usuario esté autorizado para ejecutar la consulta y compila el texto SQL del usuario en un plan de consulta interno. Una vez compilado, el plan de consulta resultante es manejado por el ejecutor del plan. El plan de ejecución consiste en un conjunto de “operadores” para ejecutar cualquier consulta.
4. En base al plan de consulta se realizan llamadas al **Transactional Storage Manager**.
5. Los métodos de acceso devuelven el control a los operadores del ejecutor de consultas, que orquestan el cálculo de las tuplas de resultado a partir de los datos de la base de datos; a medida que se generan las tuplas de resultado, se colocan en un búfer para el gestor de comunicaciones del cliente, que envía los resultados de vuelta al llamante.



Joseph M. Hellerstein, Michael Stonebraker, and James Hamilton
“ Architecture of a Database System”

Evolución de los DBMS - Michael Stonebraker

Fin del "One Size Fits Al": Hasta los 2000s, los DBMS basados en filas dominaban el mercado. Cuatro cambios clave han transformado el panorama.

- **Column Stores vs. Row Stores**: CS son superiores para data warehouses. Consultas selectivas hacen que los column stores sean 50-100 veces más rápidos.
- **DBMS en Memoria**: Caída de precios de la memoria permite bases OLTP en RAM. Arquitecturas tradicionales basadas en disco no son competitivas. Concurrencia y recuperación ante fallos requieren nuevas estrategias. MVCC y control de concurrencia determinista reemplazan el bloqueo en dos fases.
- El surgimiento del movimiento **NoSQL**
- **Hadoop** es un framework open source que permite procesar grandes volúmenes de datos de manera distribuida. Su sistema de archivos, **HDFS** (Hadoop Distributed File System), divide los datos en bloques y los distribuye entre varios nodos de un clúster. **Apache Spark** es un motor de procesamiento de datos en tiempo real y por lotes, que se integra con Hadoop y HDFS

424 systems in ranking, March 2025

Rank Mar 2025	Rank Feb 2025	Rank Mar 2024	DBMS	Database Model	Score Mar 2025	Score Feb 2025	Score Mar 2024
1.	1.	1.	Oracle	Relational, Multi-model	1253.08	-1.74	+32.02
2.	2.	2.	MySQL	Relational, Multi-model	988.13	-11.86	-113.37
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	788.14	+1.27	-57.67
4.	4.	4.	PostgreSQL	Relational, Multi-model	663.42	+3.81	+28.52
5.	5.	5.	MongoDB	Document, Multi-model	396.42	-0.21	-28.11
6.	↑ 7.	↑ 9.	Snowflake	Relational	161.78	+6.20	+36.40
7.	↓ 6.	↓ 6.	Redis	Key-value, Multi-model	155.36	-2.55	-1.64
8.	8.	↓ 7.	Elasticsearch	Multi-model	131.38	-3.25	-3.41
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model	126.57	+1.14	-1.18
10.	10.	10.	SQLite	Relational	113.08	-0.74	-5.08
11.	11.	↑ 12.	Apache Cassandra	Wide column, Multi-model	106.65	+4.07	+2.07

Ranking Marzo 2025