

Data Management and Systems (DMS) Introduction and Overview

Prof. Dr. Matthias Boehm

Technische Universität Berlin

Berlin Institute for the Foundations of Learning and Data

Big Data Engineering (DAMS Lab)

Announcements / Org



■ #1 Hybrid & Video Recording

- Hybrid lectures (in-person, zoom) with optional attendance

<https://tu-berlin.zoom.us/j/9529634787?pwd=R1ZsN1M3SC9BOU1OcFdmem9zT202UT09>

- Zoom **video recordings**, links from website

https://mboehm7.github.io/teaching/ws2627_dms/index.htm



■ #2 Course Registrations

- TU Berlin ISIS registrations as of Oct 13
- Bachelor CS/ISM vs others ratio?

~230

■ #3 Hackathon on **AI in Medicine**

- BIFOLD/Charité hackathon in planning
- Help us find suitable dates that work for you



#4 Faculty IV - Team Awareness and Antidiscrimination

<https://www.tu.berlin/eecs/awan>



- Goal: **Low-barrier approachability** for spectrum of awareness and antidiscrimination issues

- Team

- Irene Hube-Achter (MTSV)
- Matthias Boehm (professors)
- Ceenu George (professors)
- Nadine Karsten (scientific staff)
- Maria Fleßner-Jung (mentoring)
- N.N. (students)



- Mission Statement

- Account for heterogeneity and complexity of modern societies at TU Berlin
- #1 **Treat all persons with fairness and respect**
- #2 **Ensure a safe environment for all**
- #3 **Comply with our duty of care towards others**
- #4 **Actively support the implementation of the above guidelines and contribute**

Contact: private email,
eeecs-TB-awareness@win.tu-berlin.de,
or AwAn@dams.tu-berlin.de

Agenda

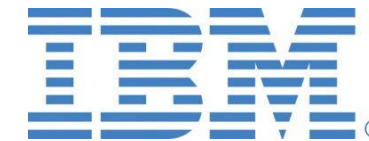


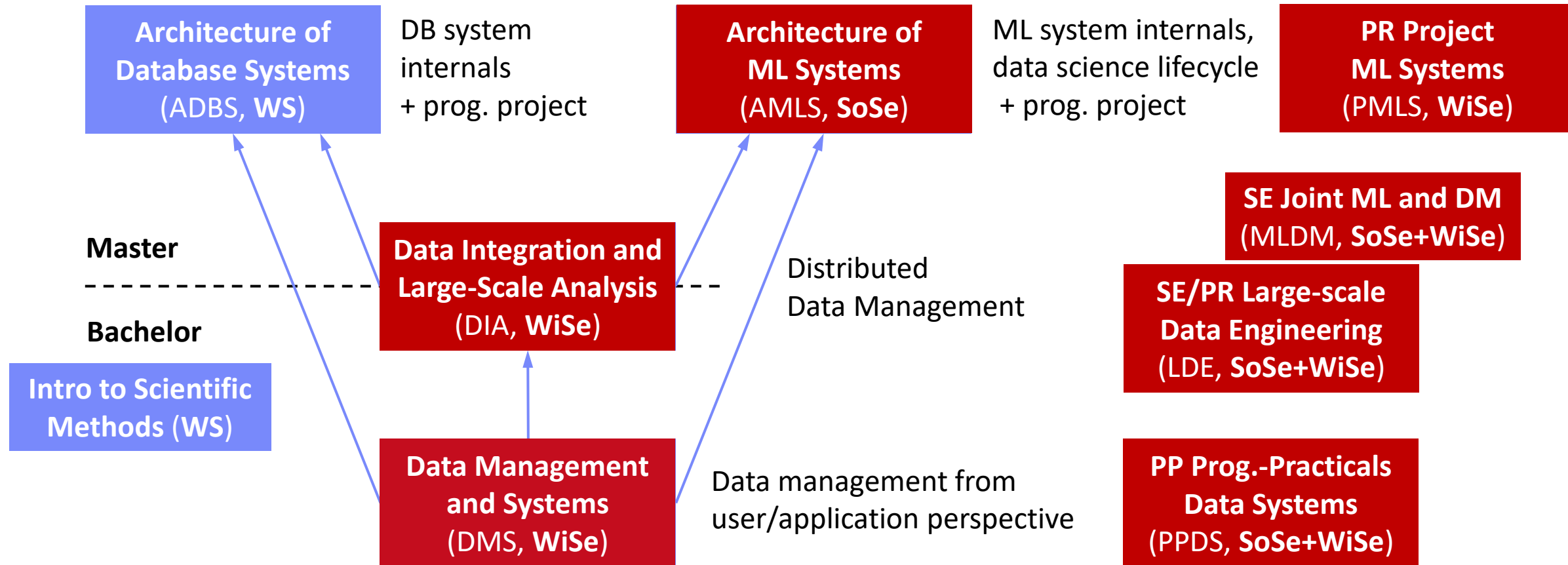
- **Data Management Group(s)**
- **Course Motivation, Goals, and Outline**
- **Course Organization and Logistics**
- **History of Data Management**

Data Management Group(s)

About Me

- **Since 09/2022 TU Berlin, Germany**
 - University professor for Big Data Engineering (DAMS)
- **2018-2022 TU Graz, Austria**
 - BMK endowed chair for data management + research area manager
 - **Data management for data science** (DAMS), **SystemDS & DAPHNE**
- **2012-2018 IBM Research – Almaden, CA, USA**
 - Declarative large-scale machine learning
 - Optimizer and runtime of **Apache SystemML**
- **2007-2011 PhD TU Dresden, Germany**
 - Cost-based optimization of integration flows
 - Time series forecasting / in-memory indexing & query processing





BIFOLD Data Management Groups



- **DIMA** (Database Systems and Information Management)
Prof Dr. **Volker Markl**
- **DAMS** (Big Data Engineering)
Prof. Dr. **Matthias Boehm**
- **D2IP** (Information Integration and Data Preparation)
Prof. Dr. **Ziawach Abedjan**
- **DEEM** (Data Engineering for Machine Learning)
Prof. Dr. **Sebastian Schelter**
- **RSiM** (Big Data Analytics for Earth Observation)
Prof. Dr. **Begüm Demir**
- **IoT** (Distributed Stream Processing in Heterogeneous Environments)
Dr. **Steffen Zeuch**

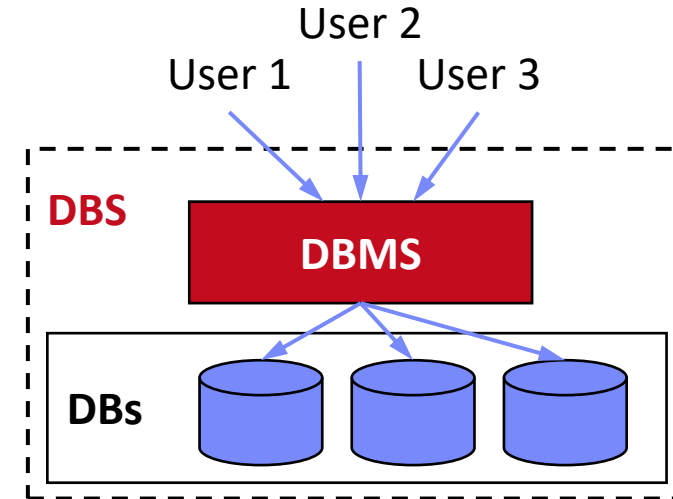


Course Motivation, Goals, and Outline

Definition and Impact

■ Def: Database System

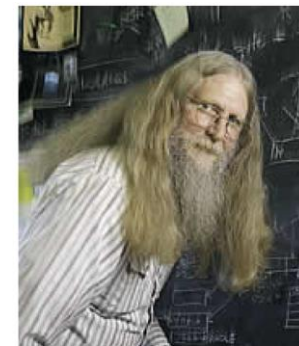
- Overall system of DBMS + DBs
- DBMS: Database Management System (SW to handle DBs)
- DBs: Database (data/metadata collection of conceptual mini-world)
- Note: DB also a short for DBS/DBMS
- Data System: broader term for database systems



■ Importance in Practice

- Market Volume: **100-250B \$US**
- Foundation of many applications in various domains

“Relational databases are the foundation of western civilization”



Bruce Lindsay



[Marianne Winslett: Bruce Lindsay speaks out: [...]. SIGMOD Record 34(2), 2005]

- **Application development and maintenance costs**
 - **Declarative queries** (what not how) and **data independence**
 - Efficient, correct, and independent data organization, size, access
- **Multi-user operations and access control**
 - Synchronization of **concurrent user queries and updates**
 - Enforce access control (e.g., permissions on tables, views)
- **Consistency and data integrity**
 - Eliminates **redundancy** and thus, enforces consistency
 - Enforces **integrity constraints** (e.g., semantic rules)
- **Logging and Recovery**
 - Recovery of consistent state after HW or SW failure
- **Performance and Scalability**
 - **High performance** for large datasets (with low resource req.) or high transaction throughput

■ Course Goals

- **A:** Understanding of **database systems** (from user perspective)
- **B:** Understanding of **modern data management** (from user perspective)
- **C:** Insights into internals of data systems → deeper understanding of algorithms & data structures, computer organization & distributed systems, foundations of programming paradigms, and logic & discrete structures

■ Meta Goals

- **Understand, use, debug, and evaluate** data management systems
- Awareness of **system alternatives** and their **tradeoffs**
- Fundamental concepts as basis for advanced courses and other areas

➔ **How: (1) from basics to modern data systems, (2) practical skills, (3) breadth and depth**

Outline: Part A (Relational) Database Systems



- **01 Introduction and Overview** [Oct 13]
- **02 Data Modeling I** [Oct 20]
- **03 Data Modeling II and Normalization** [Oct 27]
- **04 Relational Algebra** [Nov 03]
- **05 Query Languages I (SQL, XML, JSON)** [Nov 10]
- **06 Query Languages II and APIs** [Nov 17]
- **07 Physical Design and Tuning** [Nov 24]
- **08 Query Processing and Optimization** [Dec 01]
- **09 Transaction Processing and Concurrency** [Dec 08]

→ Exercise 1:
Database Creation
[Nov 11 – Dec 01]

→ Exercise 2:
Queries
[Nov 25 – Dec 15]

Outline: Part B: Modern Data Management (Systems)



- **10 Arrays and Vector Processing** [Dec 15]
- **11 NoSQL (key-value, document, graph processing)** [Jan 05]
- **12 Time Series Databases and Analytics** [Jan 12]
- **13 Distributed Storage and Data Analysis** [Jan 19]
- **14 Data Stream Processing Systems** [Jan 26]
- **15 Q&A and Exam Preparation** [Feb 02]



Exercise 3:
Vector DBs
[Dec 16 – Jan 21]

Course Organization

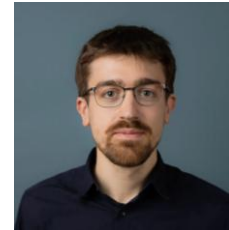
Basic Course Organization



■ Staff

- Lecturer: Prof. Dr. **Matthias Boehm** (DAMS)
- Teaching Assistants
 - **Gereon Dusella** (DIMA)
 - **David Justen** (DAMS)
 - Leonardo Dominici (DEEM)
 - Philipp Ortner (DAMS, backup)
- Student Teaching Assistants: **Vincent Johannes Kaulich**, **Sebastian Alberto Rodriguez Schmidt**

previous ISDA TAs



■ Language

- Lectures and slides: **English**
- Communication and examination: **English/German**

■ Course Format

- DMS lecture + exercise 2/2 (**3+3 ECTS**)
- **Weekly lectures** (**start 4.15pm**, including **Q&A**), **attendance optional**
- Weekly tutorials (3x, start Oct 20)

■ Communication

- **Informal language** (first name is fine)
- Please, **immediate feedback** (unclear content, missing background)
- **ISIS Q&A forum**, and email for private matters
- **Office hour**: separately announced before exercise submission

■ Exercises

- Written and programming exercises, submitted/graded through **ISIS**; completed if **>75% each**
- Deadlines are important (**at most 7 late days in total**)
- Individual assignments (**academic honesty / no plagiarism**, incl GenAI)



■ Website

- https://mboehm7.github.io/teaching/ws2627_dms/index.htm



■ Exams

- Feb 22, 2026, 3pm (first) and Mar 24, 2026, 2pm (second); laptop in-presence w/o private devices/materials

Timetable Lectures / Tutorials

For tutorials: sign-up for the tutorials in MOSES by Oct 14 6pm



	Mon	Tue	Wed	Thu	Fri
08:00					
09:00					
10:00					
11:00					
12:00			Datenmanagement und Datensystem Grossgruppenübung 01 Charlottenburg, C 243		Datenmanagement und Datensystem Grossgruppenübung 03 Charlottenburg, EW 203
13:00					
14:00					
15:00					
16:00		Datenmanagement und Datensystem Vorlesung Charlottenburg, A 053 Böhm, Matthias		Datenmanagement und Datensystem Grossgruppenübung 02 Charlottenburg, ER 164	
17:00					



- **Not needed** for lectures / exercises (course is self-contained), **but second perspective**
- Raghu Ramakrishnan, Johannes Gehrke: Database Management Systems (3. ed.). McGraw-Hill 2003, ISBN 978-0-07-115110-8, pp. I-XXXII, 1-1065
- Jeffrey D. Ullman, Jennifer Widom: A first course in database systems (2. ed.). Prentice Hall 2002, ISBN 978-0-13-035300-9, pp. I-XVI, 1-511
- Ramez Elmasri, Shamkant B. Navathe: Fundamentals of Database Systems, 3rd Edition. Addison-Wesley-Longman 2000, ISBN 978-0-8053-1755-8, pp. I-XXVII, 1-955
- Alfons Kemper, André Eickler: Datenbanksysteme - Eine Einführung, 10. Auflage. De Gruyter Studium, de Gruyter Oldenbourg 2015, ISBN 978-3-11-044375-2, pp. 1-879

Additional Perspective: [Zachary G. Ives, Rachel Pottinger, Arun Kumar, Johannes Gehrke, Jana Giceva: The future of data(base) education: Is the "cow book" dead?, VLDB 2021]



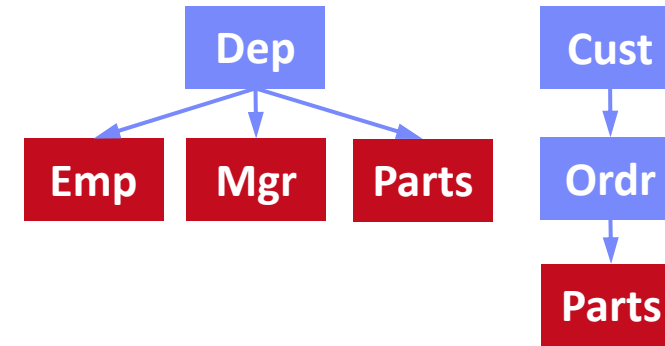
History of Data Management

History 1960/70s (pre-relational)



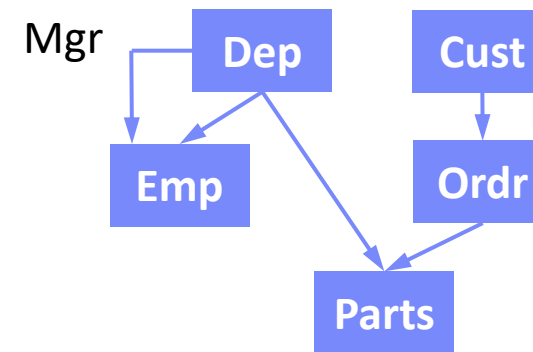
■ Hierarchical Model

- Tree of records
- E.g., IBM Information Management System (IMS) – IMS 15 (Oct 2017)



■ Network Model

- CODASYL (COBOL, DB interfaces)
- Graph of records
- Charles Bachman ([Turing Award '73](#))
- E.g., Integrated Data Store (IDS)



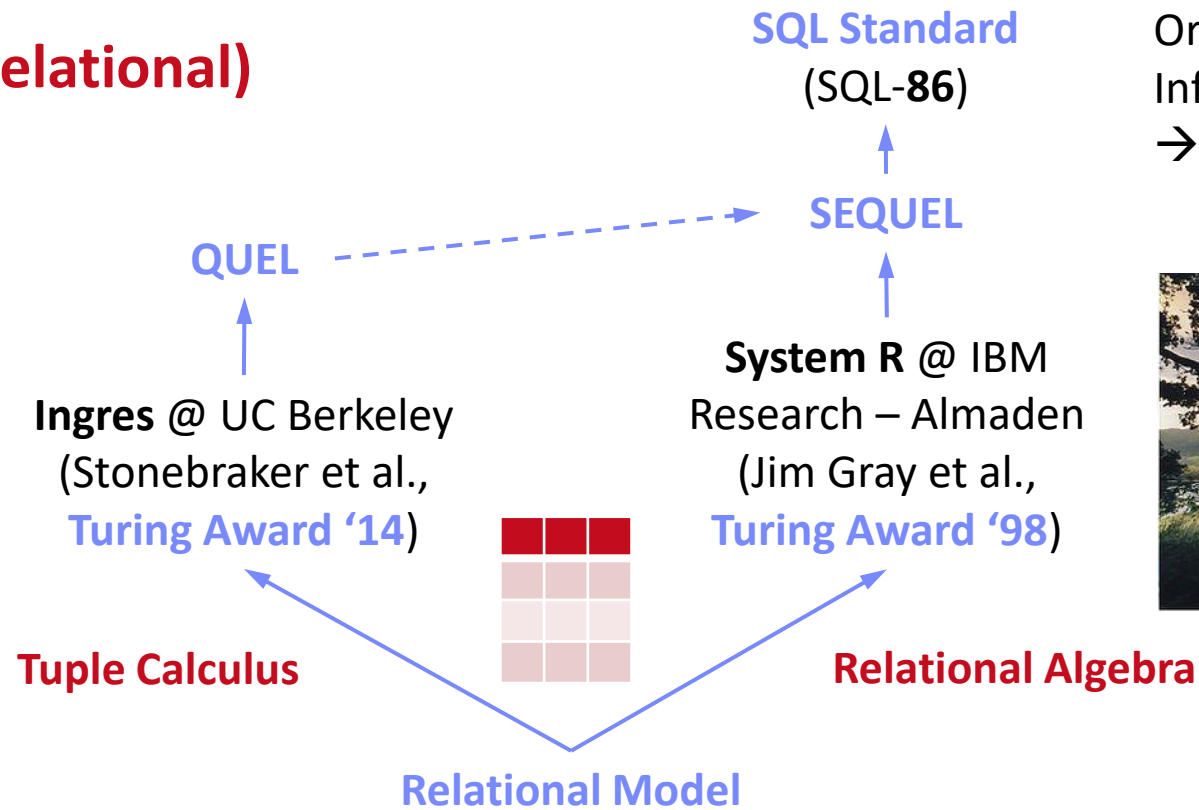
CODASYL: Conference on Data Systems Languages
COBOL: Common Business Oriented Language

■ Pros and Cons (see [NoSQL Doc-Stores](#))

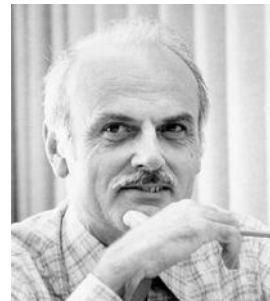
- Performance by **directly traversing static links**
- **Duplicates** → inconsistencies on updates, **data dependence**

History 1970/1980s (relational)

Oracle, IBM DB2,
Informix, Sybase
→ MS SQL



- Goal: Data Independence**
(physical data independence)
- Ordering Dependence
 - Indexing Dependence
 - Access Path Dependence



Edgar F. "Ted" Codd @ IBM
Research (**Turing Award '81**)

[E. F. Codd: A Relational Model of
Data for Large Shared Data Banks.
Comm. ACM 13(6), **1970**]



Success of SQL / Relational Model



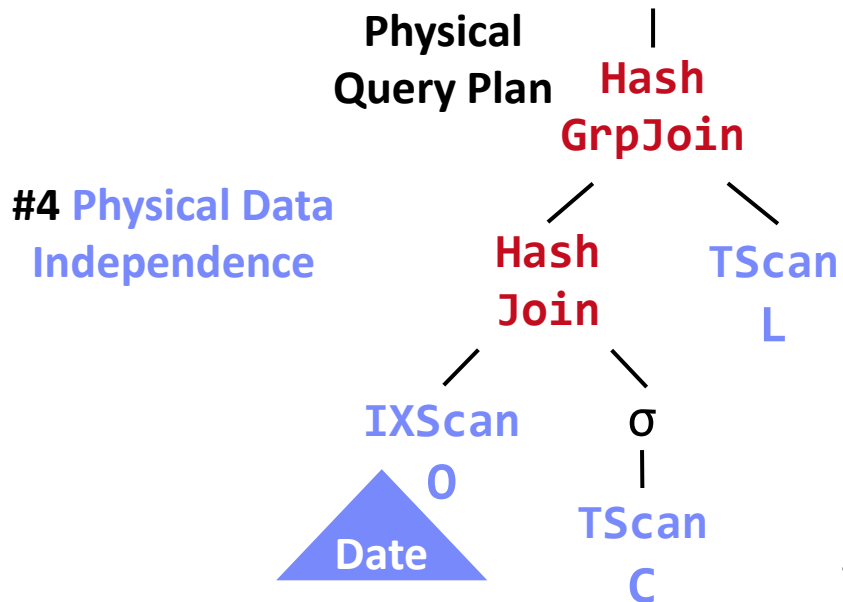
#1 **Declarative:**
what not how

#2 **Flexibility:**
closure property
→ composability

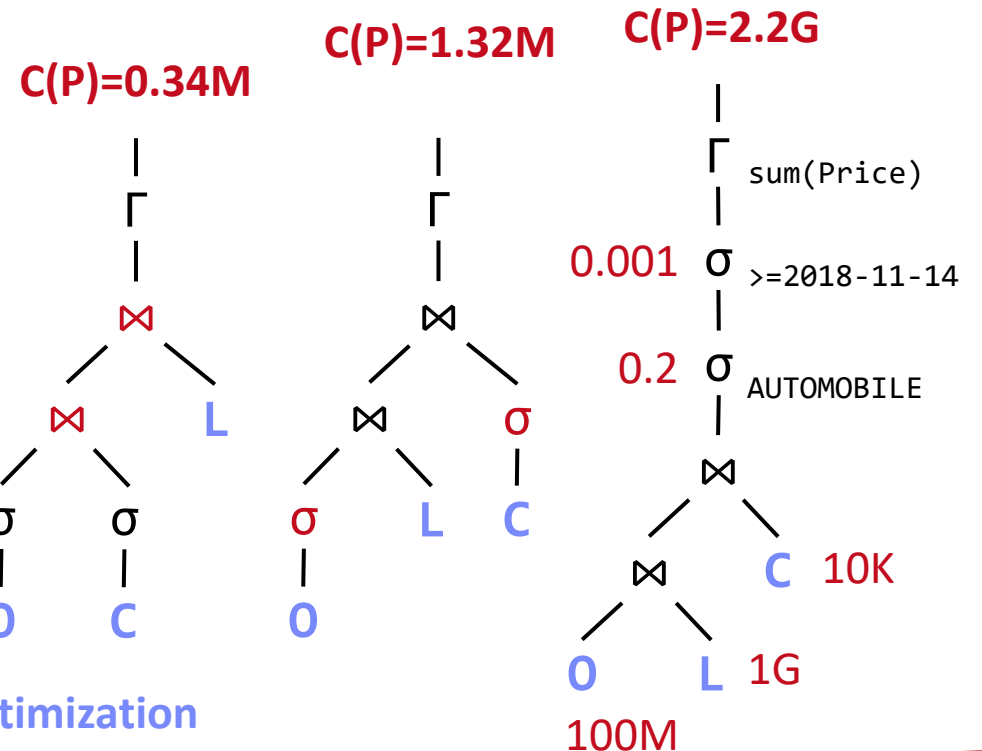
Query:

```
SELECT O_OID, sum(L_Price)
FROM Orders, Lineitem, Customer
WHERE O_OID = L_OID AND O_CID = C_CID
      AND O_Odate >= '2018-11-14'
      AND C_Msegment = 'AUTOMOBILE'
GROUP BY O_OID
```

Logical Query Plans



#3 **Automatic Optimization**



Excursus: PostgreSQL



- **History of PostgreSQL** (used in examples/exercises)
 - Postgres is the successor project of commercialized Ingres
 - Focus on abstract data types, commercialized as Illustra
 - Prototype w/ SQL open sourced as **Postgres95** → **PostgreSQL**
 - Heavily used as **basis for research projects / startups**
- **Recommended Reading**
 - Michael Stonebraker: The land sharks are on the squawk box. Commun. ACM 59(2): 74-83 (2016), Turing Award Lecture, <https://dl.acm.org/citation.cfm?doid=2886013.2869958>
 - **Video:** <http://www.youtube.com/watch?v=sEPTZVGk3WY>
 - **Slides:** <http://vldb.org/2015/wp-content/uploads/2015/09/stonebraker.pdf>



History 1980/90/2000s

OLTP ... Online Transaction Processing
OLAP ... Online Analytical Processing
ETL ... Extract, Transform, Load



■ Enterprise DBMS

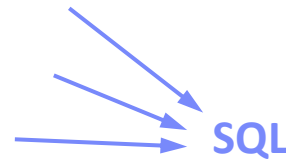
- Heavy investment in research and development → adoption
- Oracle, IBM DB2, Informix, Sybase, MS SQL, PostgreSQL, MySQL
- Other technologies: OODBMS, Multimedia, Spatiotemporal, Web, XML

■ Information/Data Warehousing (DWH)

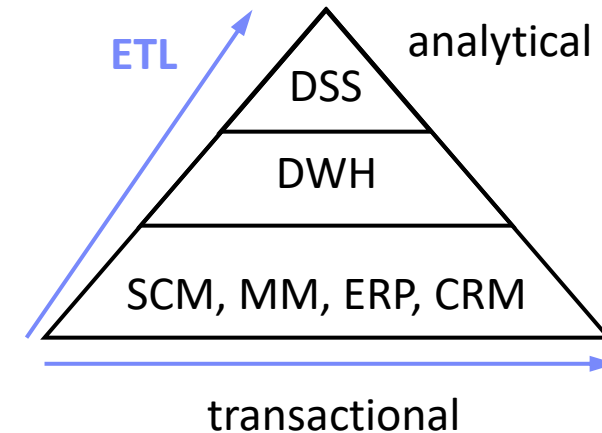
- Workload separation into OLTP and OLAP
- Classical DWH architecture: operational, staging, DWH, data marts + mining
- ETL Process (Extract, Transform, Load)

■ Different Personas

- Domain Experts (e.g., BI Tools, SAP R/3)
- DB Application Developers (e.g., ABAP)
- DB Developers and DB Admins



SQL



■ Specialized Systems

- Column stores + compression for OLAP
- Main memory systems for OLTP and OLAP
- Data streaming, scientific and graph databases
- Information extraction / retrieval, and XML

■ Other Research Trends

- Approximate QP / Adaptive QP / tuning tools
- Large-scale data management (DFS, MR) / cloud computing

■ Toward Flexible, Large-Scale Data Management

(DWH ... a bygone era)

- MAD Skills (magnetic, agile, deep), MADlib
- Integration of R, Python in data analysis
- Open data and its integration, query processing over raw (open) data formats

[M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem, P. Helland: **The End of an Architectural Era** (It's Time for a Complete Rewrite). **VLDB 2007**]



[J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, C. Welton: **MAD Skills**: New Analysis Practices for Big Data. **PVLDB 2(2) 2009**]

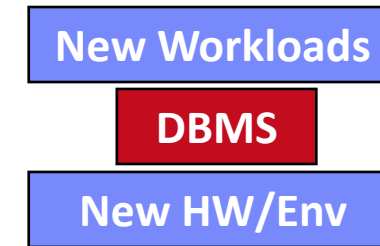


History 2010s – Present



■ Two Key Drivers of DB Research

- **New analysis workloads** (NLP, key/value, RDF/graphs, documents, time series, ML) and applications
- **New HW/infrastructure** (multi-/many-core, cloud, scale-up/scale-out, NUMA/HBM, RDMA, SSD/NVM, FPGA/GPU/ASIC)

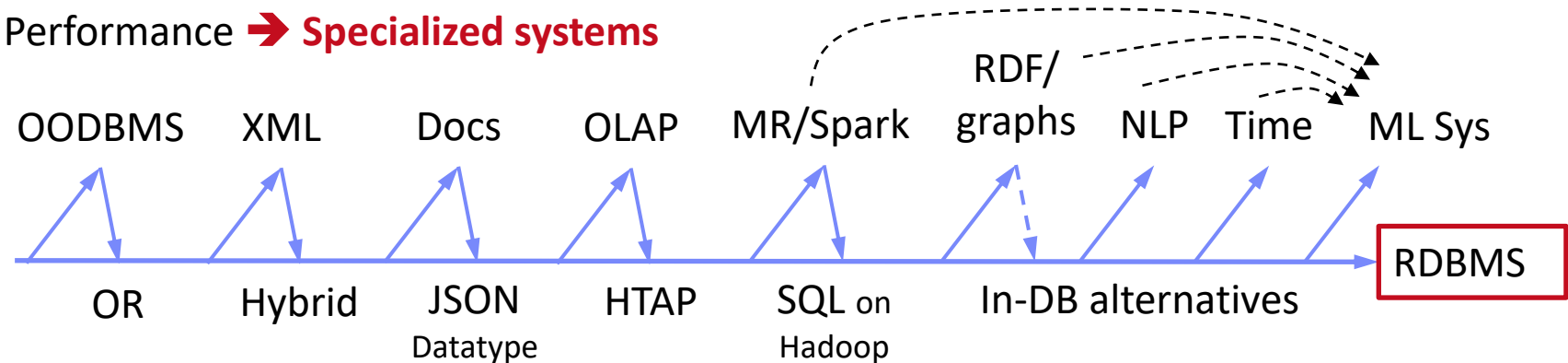


■ Excursus: A retrospective view of specialized systems

- **Goal #1:** Avoid boundary crossing → **General-purpose**
- **Goal #2:** New workload + Performance → **Specialized systems**
- Some Examples



[Michael Stonebraker, Andrew Pavlo: What Goes Around Comes Around... And Around... **SIGMOD Rec. 2024.**]



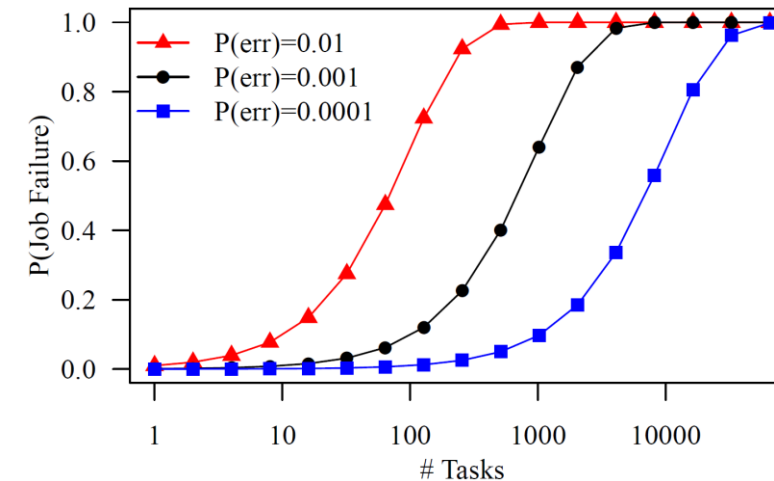
- **Motivation NoSQL Systems**
 - **Flexible schema** (no upfront costs), **scalability**, or specific data types
 - Relaxed **ACID** (atomicity, consistency, isolation, durability) requirements
 - ➔ **BASE** (basically available, soft state, eventual consistency)
- **Example NoSQL Systems** (local and distributed):
 - **Key/Value-Stores**: simple put/get/delete, massive scalability
 - **Document-Stores**: store nested documents (tree)
 - **RDF Stores**: store subject-predicate-object triples
 - **Graph DBs**: store nodes/edges/attributes, vertex-centric
 - **Time Series DBs**: store sequences of observations
 - **Vector DBs**: store vector embeddings of different modalities, similarity search

History 2010s – Present, cont.



■ Motivation Large-Scale Data Management

- **Massive scalability** (data/compute) on demand, cost effective **fault tolerance**, flexibility
- Example Facebook 2014:
300PB DWH, **600TB** daily ingest
- Cost-effective commodity hardware
- Error rate increases with increasing scale

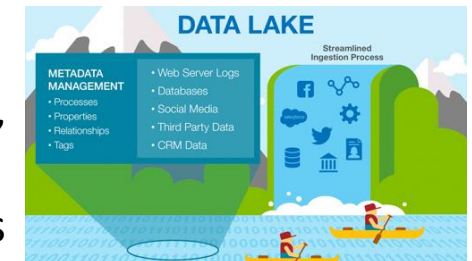


■ Examples Large-Scale Data Management

- Distributed file systems w/ replication (e.g., GPFS, HDFS)
- **Cloud object storage** (e.g., Amazon s3, OpenStack Swift)
- Data-parallel data analysis with Spark/Flink, incl streaming
- Automatic cloud resource elasticity (pay as you go)

} **Data Lake**

Audio, Image,
Video, Text,
Streams, Logs



History 2020s – Present



■ New Trends w/ Large Language Models (LLMs), GenAI, and Coding Agents

[Generated with Gemini 3.1 Pro]

[Olga Ovcharenko, Matthias Boehm, Sebastian Schelter: **SemPipes** - Optimizable Semantic Data Operators for Tabular Machine Learning Pipelines. **arxiv**, under submission 2026]



SPECTRUM OF MODERN QUERY & DATABASE OPTIMIZATION TECHNIQUES

HUMAN-CENTRIC & DECLARATIVE

SYSTEM-CENTRIC & OPTIMIZED

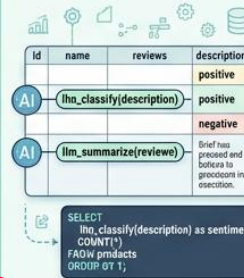
1. TEXT-TO-SQL



Focus: Natural Language to Structured Query.

- Low-code / No-code interfaces
- Intent-based exploration
- Requires semantic mapping.

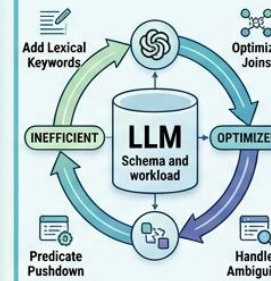
2. SEMANTIC OPERATORS



Focus: SQL-Native AI Reasoning on Data.

- Composability
- ACID compliant LLM calls
- Reduces data movement.

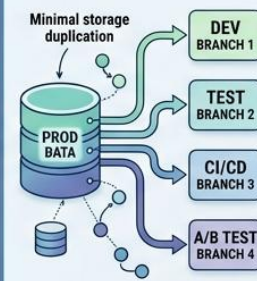
3. LLM-ASSISTED QUERY REWRITING



Focus: Automatic Query Tuning before Execution.

- Workload-aware optimization
- Iterative refinement
- Statistical accuracy guarantees.

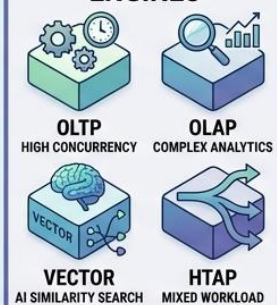
4. BRANCHING-AWARE DBMS



Focus: Isolated, Rapid Data Environments.

- Copy-on-write storage
- Instant cloning
- Supports Git-like workflows.

5. WORKLOAD-OPTIMIZED DB ENGINES



Focus: Tailored Performance for Specific Tasks.

- Columnar vs. Row storage
- Specialized index structures
- Resource allocation.

[Saeed Fathollahzadeh, Essam Mansour, Matthias Boehm: **ReSequel**: Robust LLM-assisted Query Rewriting and Optimization using Templatzation and Sampling. **PVLDB 2026**]



[Johannes Wehrstein, Timo Eckmann, Matthias Jasny, Carsten Binnig: **Bespoke OLAP**: Synthesizing Workload-Specific One-size-fits-one Database Engines. **PVLDB 2026**]



■ #1 Database Systems

- Mature and established technology → broadly applicable & eco system
- **General concepts:** abstraction, data modeling, query optimization & processing, transaction processing and recovery, physical design and tuning

■ #2 Modern Data Management

- Multiple specialized systems for specific scale / data types
- General trend toward less upfront cost, flexibility, and higher scalability

→ Variety of data management tools → **Course meta goals**

- **Understand, use, debug**, and evaluate data management systems
- **Fundamental concepts** as basis for advanced courses and other areas

■ Next Lectures

- **02 Data Modeling I** [Oct 20] (ER Diagrams)
- **03 Data Modeling II and Normalization** [Oct 27] (ERD → Relational Model)