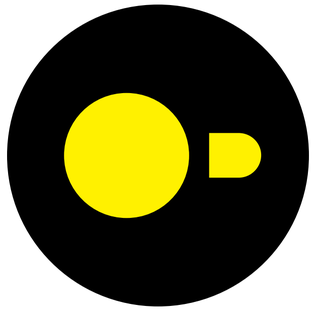


A Short Summary of the Last Decades of Data Management

Prof. Dr. Hannes Mühleisen

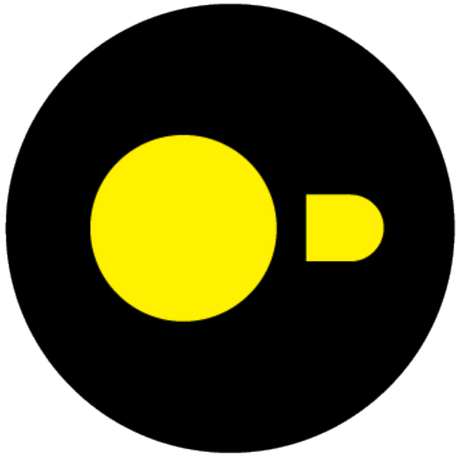
DuckDB


D
S QL 1337
BADEN-WUERTTEMBERG

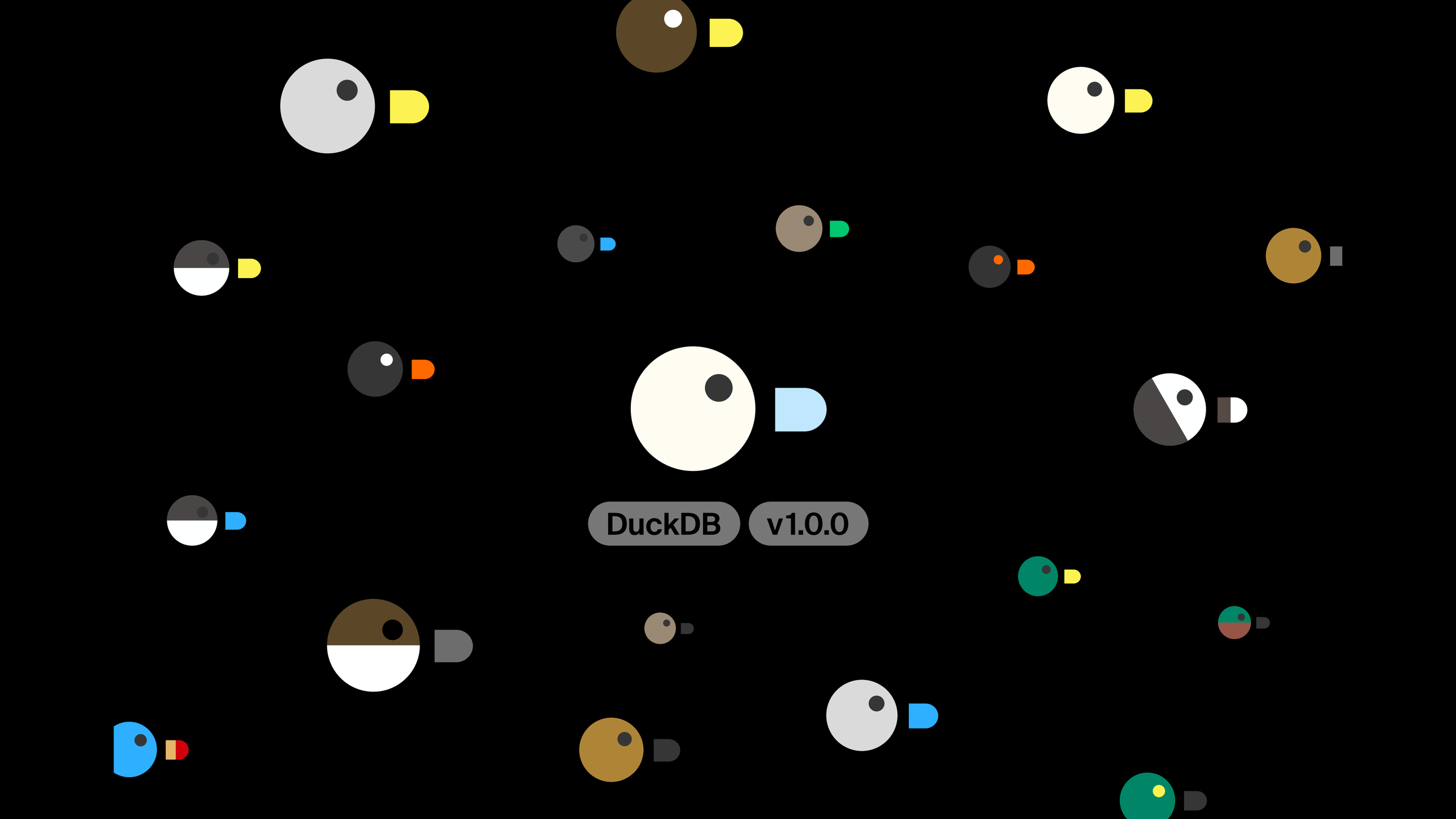
Researcher

Co-Founder & CEO

Professor



DuckDB


DuckDB
v1.0.0

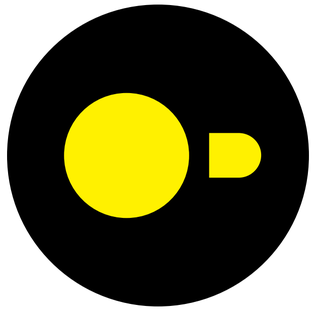
1. ▲ POV-Ray – The Persistence of Vision Raytracer (povray.org)
   114 points by RafelMri 4 hours ago | hide | 59 comments
2. ▲ macOS Sequoia to Allow iCloud Logins in Virtual Machines on ARM Macs (developer.apple.com)
   89 points by throwaway-blaze 4 hours ago | hide | 10 comments
3. ▲ Exploring Gleam, a type-safe language on the BEAM (christopher.engineering)
   92 points by crowdhailer 4 hours ago | hide | 35 comments
4. ▲ Apple's On-Device and Server Foundation Models (machinelearning.apple.com)
   712 points by 2bit 13 hours ago | hide | 362 comments
5. ▲ NanoGPT: The simplest, fastest repository for training medium-sized GPTs (github.com/karpathy)
   74 points by ulrischa 5 hours ago | hide | 8 comments
6. ▲ DuckDB Isn't Just Fast (csvbase.com)
   69 points by calpaterson 6 hours ago | hide | 21 comments
7. ▲ OpenWorm – creating a virtual organism in a computer (openworm.org)
   27 points by skilled 4 hours ago | hide | 7 comments
8. ▲ Free Quality SoundFonts (Sf2) (sites.google.com)
   66 points by hggh 6 hours ago | hide | 10 comments
9. ▲ Private Cloud Compute: A new frontier for AI privacy in the cloud (security.apple.com)
   465 points by serhack_ 13 hours ago | hide | 236 comments
10. ▲ I built an ROV to solve missing person cases (suanto.com)
    495 points by craydandy 17 hours ago | hide | 94 comments
11. ▲ The Backbone of Cybersecurity: Hardware Security Modules (join.tech)
    4 points by 5n00py 1 hour ago | hide | 1 comment
12. ▲ Engage your audience: get to the point, use story structure, force specificity (iandanielstewart.com)
    129 points by ingve 13 hours ago | hide | 34 comments
13. ▲ Big Tech's role in enabling link fraud – take 2 (eligrey.com)

Wednesday Jun 12 14:30 – 15:20
Veilingzaal

DuckDB: Crunching data anywhere, from laptops to servers

Gábor Szárnyas

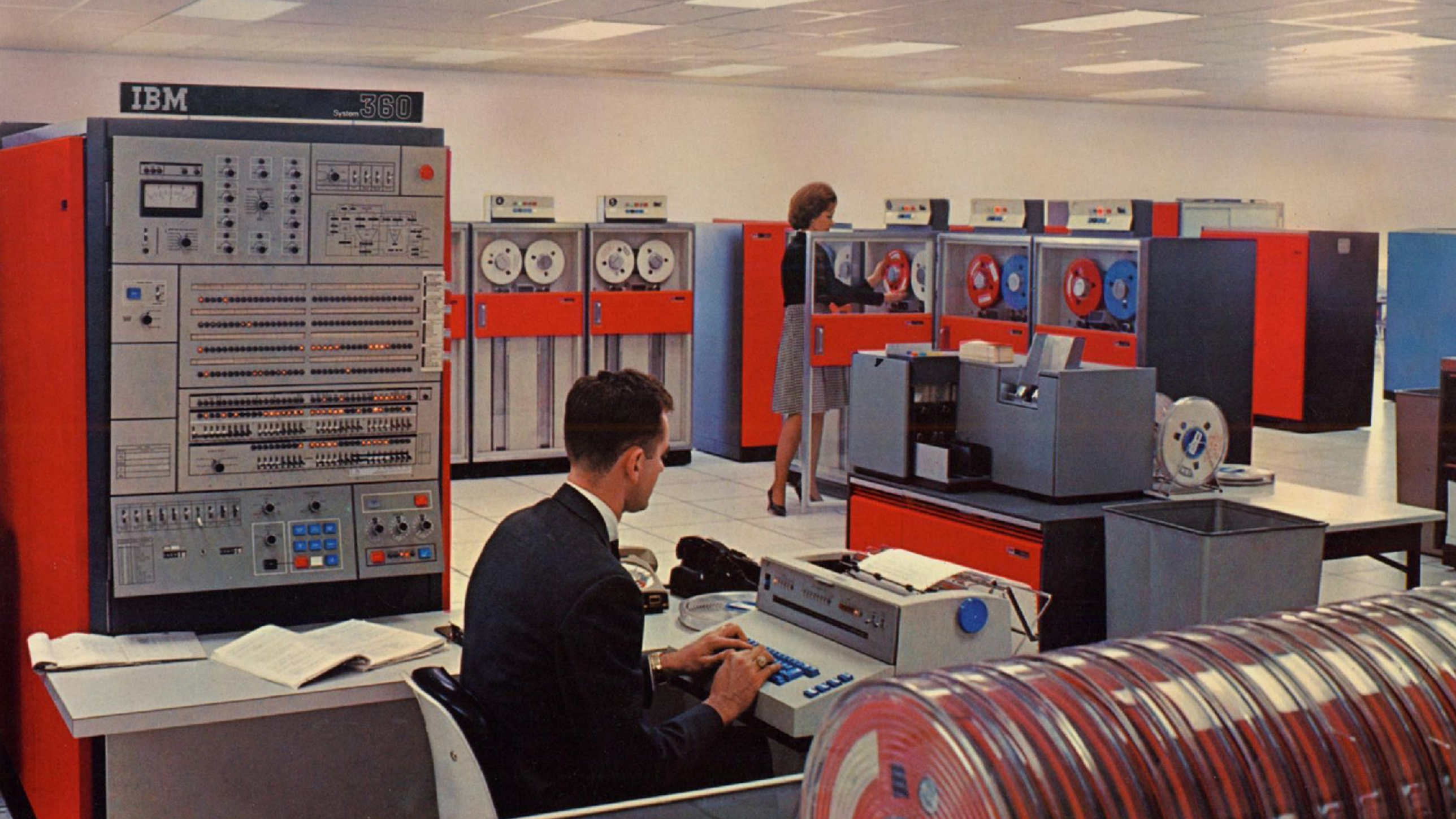
Technical Writer at DuckDB



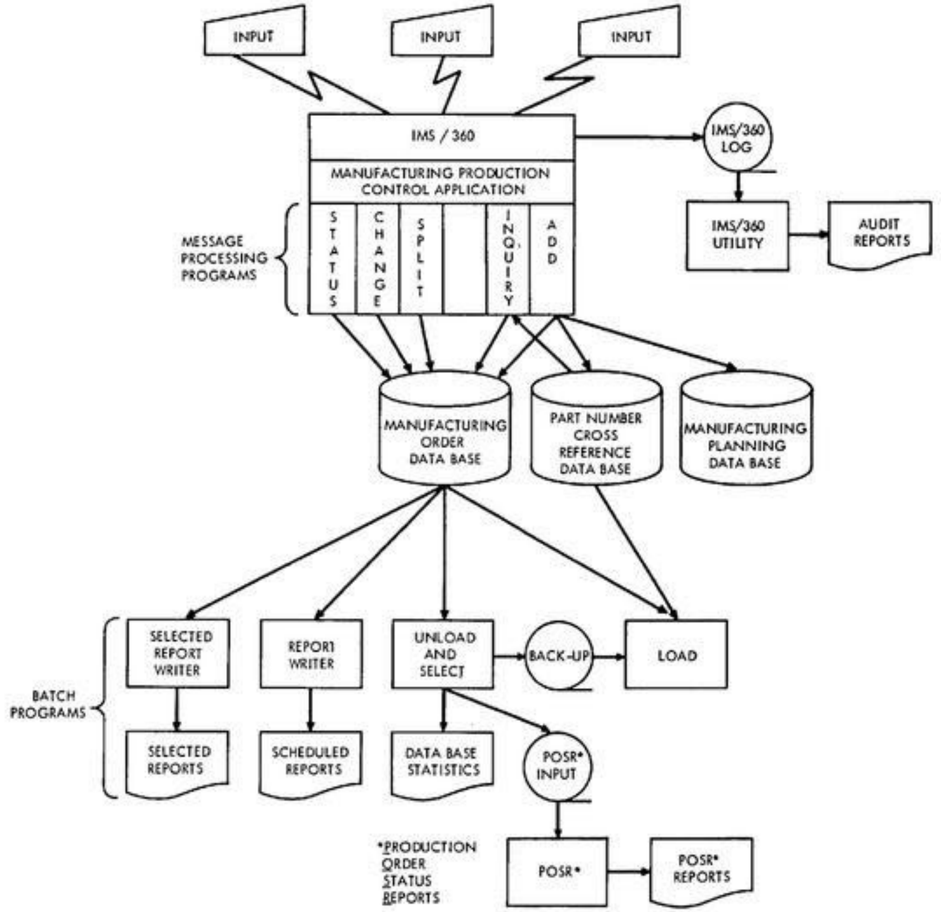
Ancient History

Item	Amount
Hop	42
Barley	84

UNITED
STATES


IBM
System 360

ICS/DL/I
IMS


INPUT
INPUT
INPUT
IMS / 360
MANUFACTURING PRODUCTION CONTROL APPLICATION
IMS/360 LOG
IMS/360 UTILITY
AUDIT REPORTS
MESSAGE PROCESSING PROGRAMS
STATUS
CHANGE
SPLIT
INQUIRY
ADD
MANUFACTURING ORDER DATA BASE
PART NUMBER CROSS REFERENCE DATA BASE
MANUFACTURING PLANNING DATA BASE
BATCH PROGRAMS
SELECTED REPORT WRITER
REPORT WRITER
UNLOAD AND SELECT
BACK-UP
LOAD
SELECTED REPORTS
SCHEDULED REPORTS
DATA BASE STATISTICS
POSR* INPUT
*PRODUCTION ORDER STATUS REPORTS
POSR*
POSR* REPORTS

A Relational Model of Data for Large Shared Data Banks

E. F. CODD
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

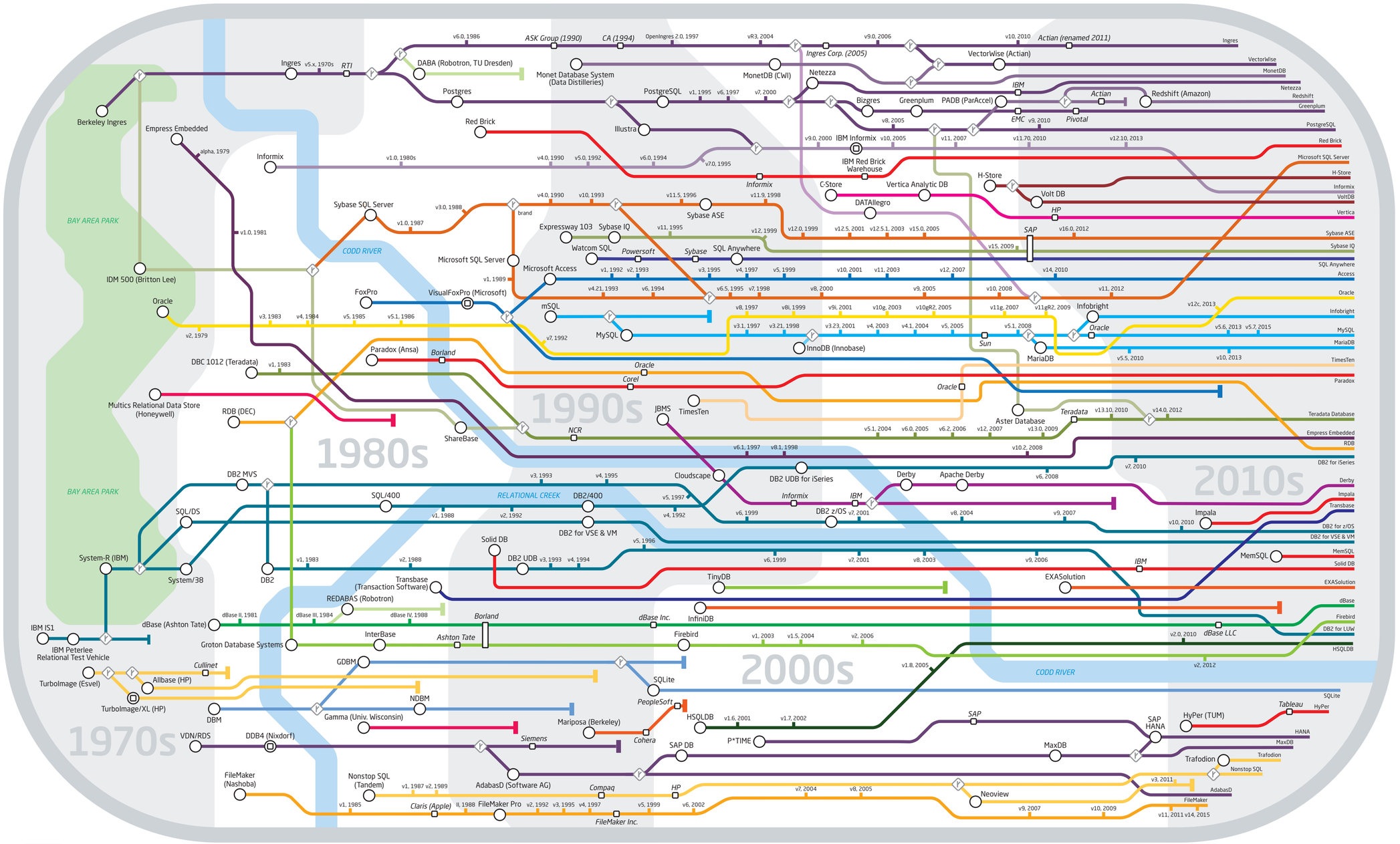
The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted

1973 IBM System R
1979 Oracle
1983 IBM DB/2
1996 PostgreSQL

Genealogy of Relational Database Management Systems



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Key to lines and symbols

DBMS name (Company) | *Acquisition* | v9, 2006 Versions | Discontinued | Branch (intellectual and/or code) | Crossing lines have no special semantics

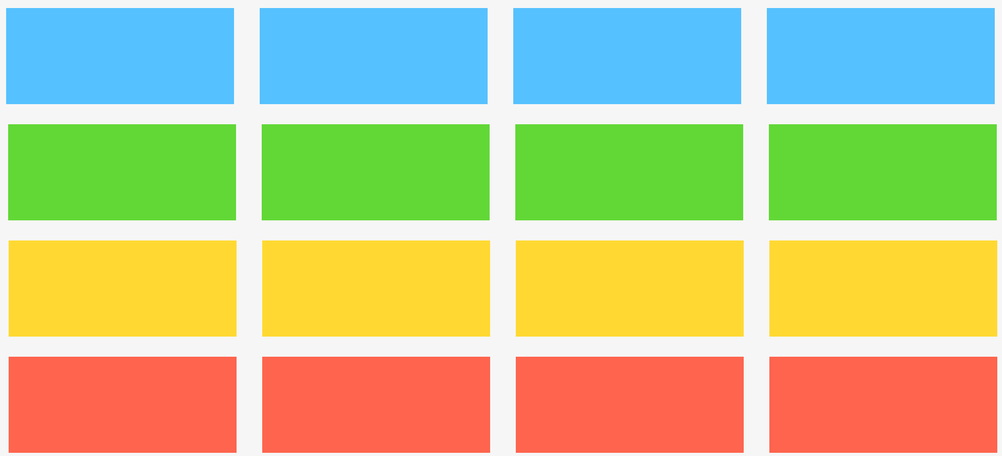
Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact - Hasso Plattner Institut, University of Potsdam, felix.naumann@hpi.de
Design - Alexander Sandt Grafik-Design, Hamburg

Version 6.0 - October 2018
https://hpi.de/naumann/projects/rdbms-genealogy.html

One Size Fits All?

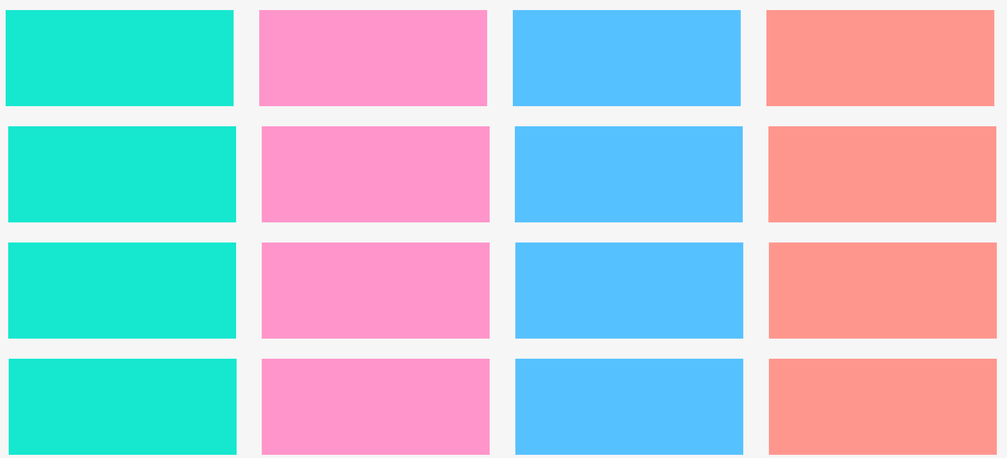
Transactional

row-based



Analytical

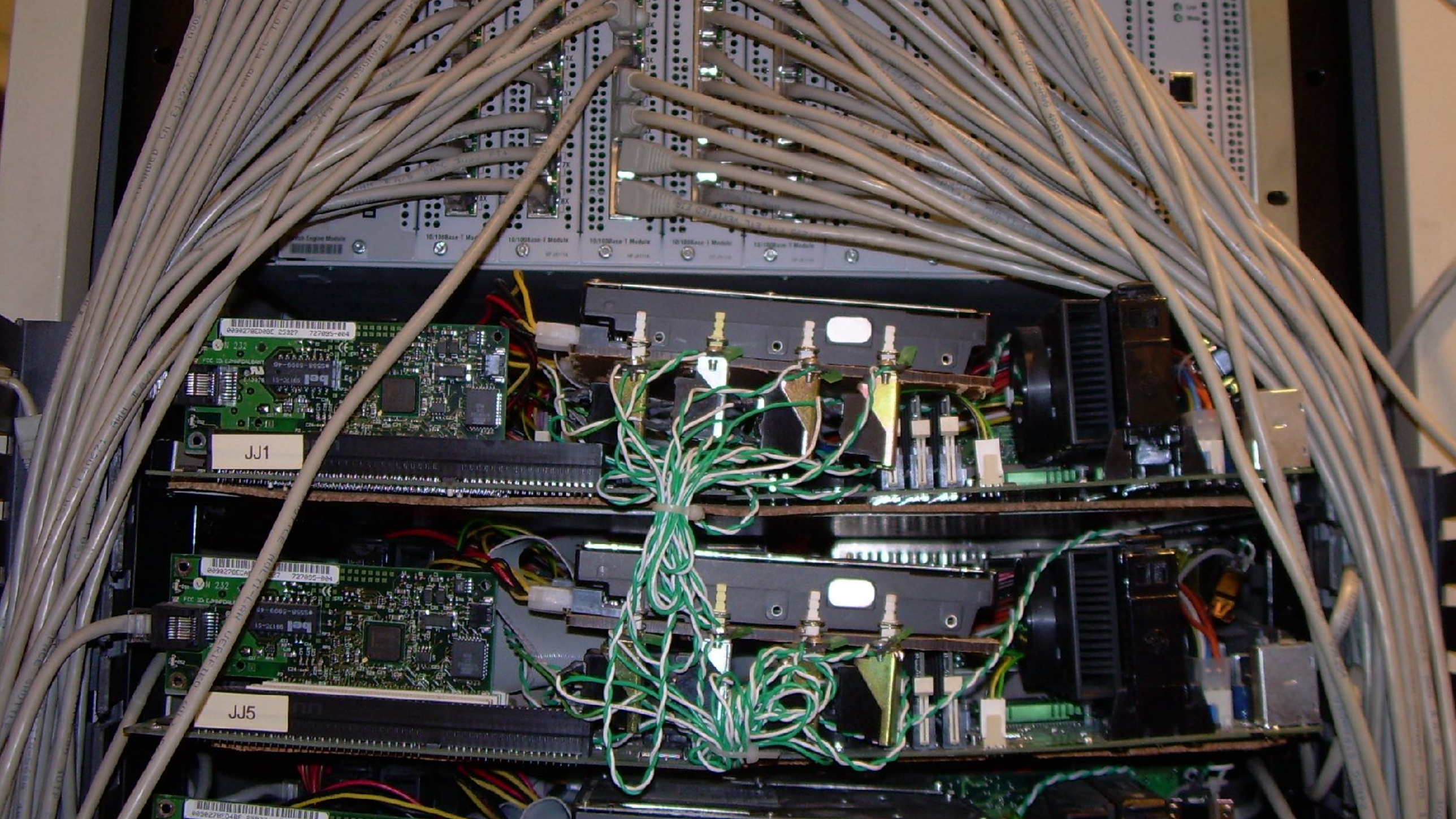
column-based



Transactional

Analytical

No SQL!


JJ1
JJ5

2006

No declarative query language.

For scale!

~~2006~~

~~No declarative query language.~~

~~For scale!~~

2010

Let’s have SQL again

2009

No schema.

For scale!

mongoDB®

~~2009~~

~~No schema.~~

~~For scale!~~

2017

Schema validation

mongoDB®

2008

No ACID transactions.

For scale!

~~2008~~

~~No ACID transactions.~~

~~For scale!~~

2023

Support ACID after all

2014

No Internal Storage

For freedom!

~~2014~~

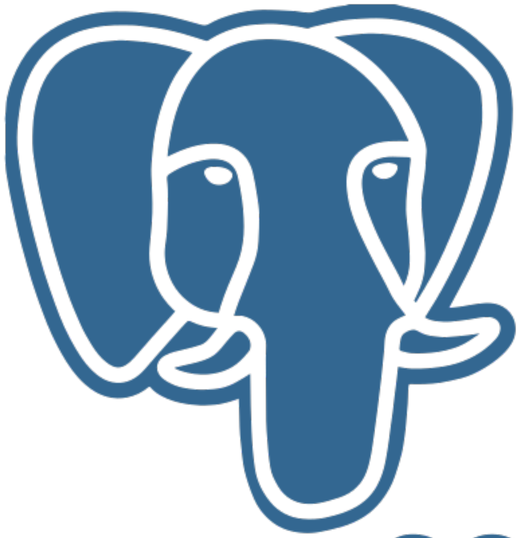
~~No Internal Storage~~

~~For freedom!~~

2024

Internal Storage

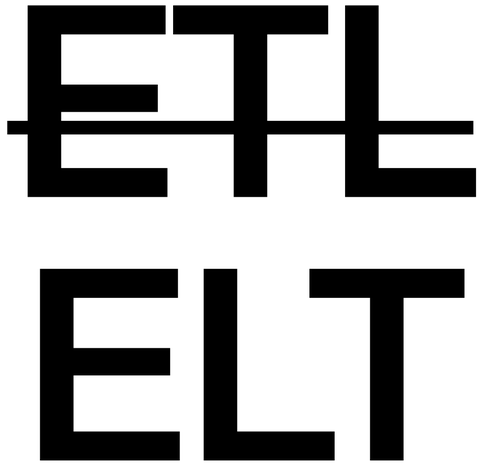
Tables
SQL
ACID

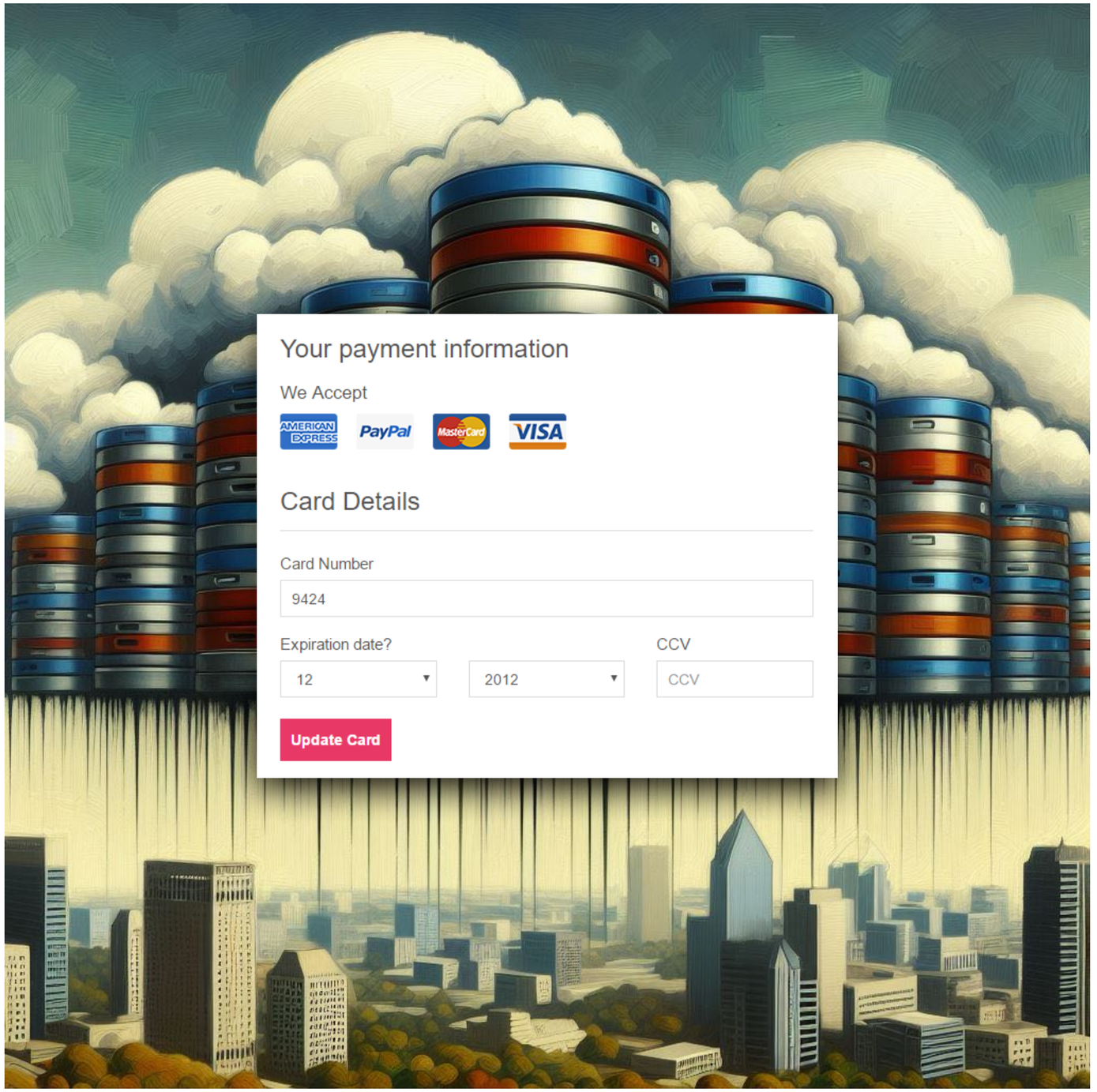


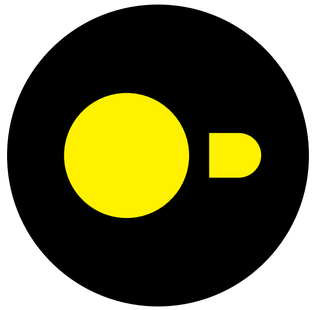
PostgreSQL


SQLite

New SQL!


ETL
ELT


Your payment information
We Accept
AMERICAN EXPRESS
PayPal
MasterCard
VISA
Card Details
Card Number
9424
Expiration date?
12
2012
CCV
CCV
Update Card



Relational eats Everything*

XML
RDF
Blockchain
ORM
...

Key/Value

Key VARCHAR	Value VARCHAR

Document

Key VARCHAR	Value JSON

Time Series

Time TIMESTAMP	Measurement DOUBLE

Vectors

Key VARCHAR	Embedding FLOAT[256]

Oracle Database Insider

All things database: the latest news, best practices, code examples, cloud, and more

Follow:

Artificial Intelligence

Oracle Announces General Availability of AI Vector Search in Oracle Database 23ai

May 2, 2024 | 5 minute read

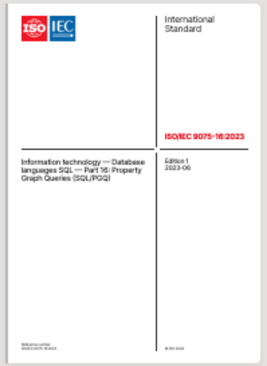

Doug Hood
Oracle AI Vector Search Product Manager


Ranjan Priyadarshi
Senior Director, Product Management(Mission-Critical Data and AI Engines)

Graphs

ID BIGINT	Properties KEY/VALUE

Source BIGINT	Target BIGINT	Properties KEY/VALUE



Read sample

ISO/IEC 9075-16:2023

Information technology — Database languages SQL

Part 16: Property Graph Queries (SQL/PGQ)

Published (Edition 1, 2023)

Data Frames

Full Text Search

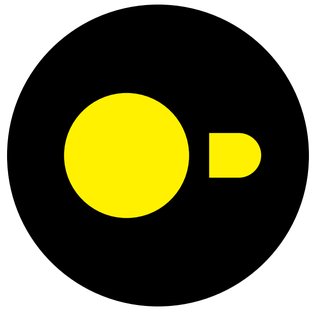
Old Dogs Are Great at New Tricks: Column Stores for IR Prototyping

Hannes Mühleisen,[1] Thaer Samar,[1] Jimmy Lin,[2] and Arjen de Vries[1]

TermID	Term
1	put
2	robe
3	wizard
4	hat

TermID	DocID	Pos
1	1	2
2	1	5
3	1	7
4	1	8

What refuses to be eaten?



Big Data Is Dead?

Human typing speed	200	Characters/s
NL working population	10 000 000	
NL work year	100 000	Minutes
Total data produced	200 000 000 000 000	Bytes
	200	Terabytes
After compression	60	Terabytes


SEAGATE
mozaic 3+
EXOS™
SEAGATE
mozaic 3+
EXOS™

Scalability! But at what COST?

Frank McSherry
Unaffiliated

Michael Isard
Unaffiliated*

Derek G. Murray
Unaffiliated†

Abstract

We offer a new metric for big data platforms, COST, or the Configuration that Outperforms a Single Thread. The COST of a given platform for a given problem is the hardware configuration required before the platform outperforms a competent single-threaded implementation. COST weighs a system's scalability against the overheads introduced by the system, and indicates the actual performance gains of the system, without rewarding systems that bring substantial but parallelizable overheads.

We survey measurements of data-parallel systems recently reported in SOSP and OSDI, and find that many systems have either a surprisingly large COST, often

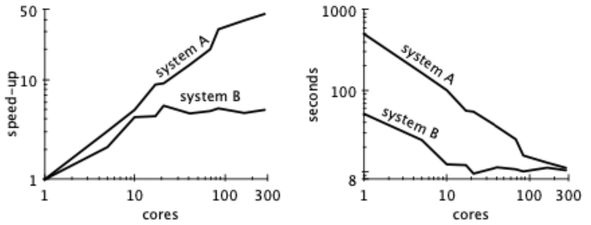
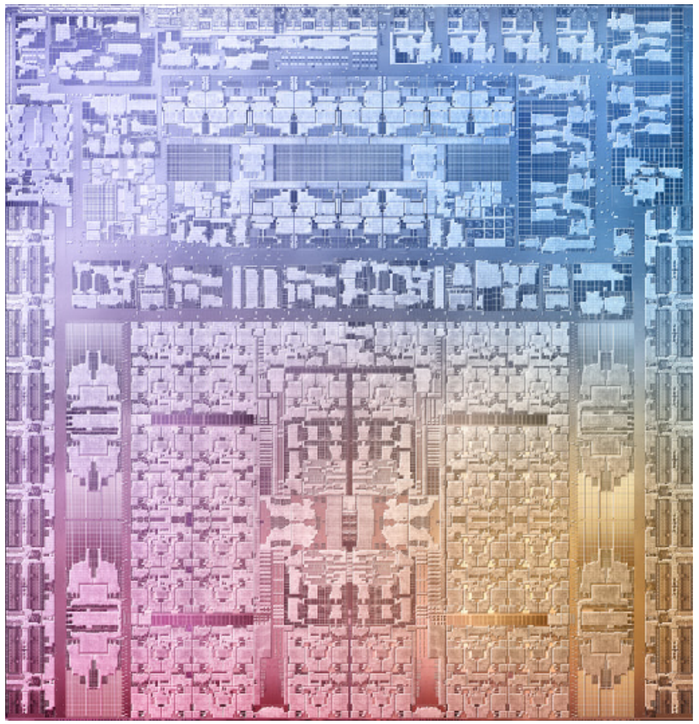


Figure 1: Scaling and performance measurements for a data-parallel algorithm, before (system A) and after (system B) a simple performance optimization. The unoptimized implementation "scales" far better, despite (or rather, because of) its poor performance.

CPUs



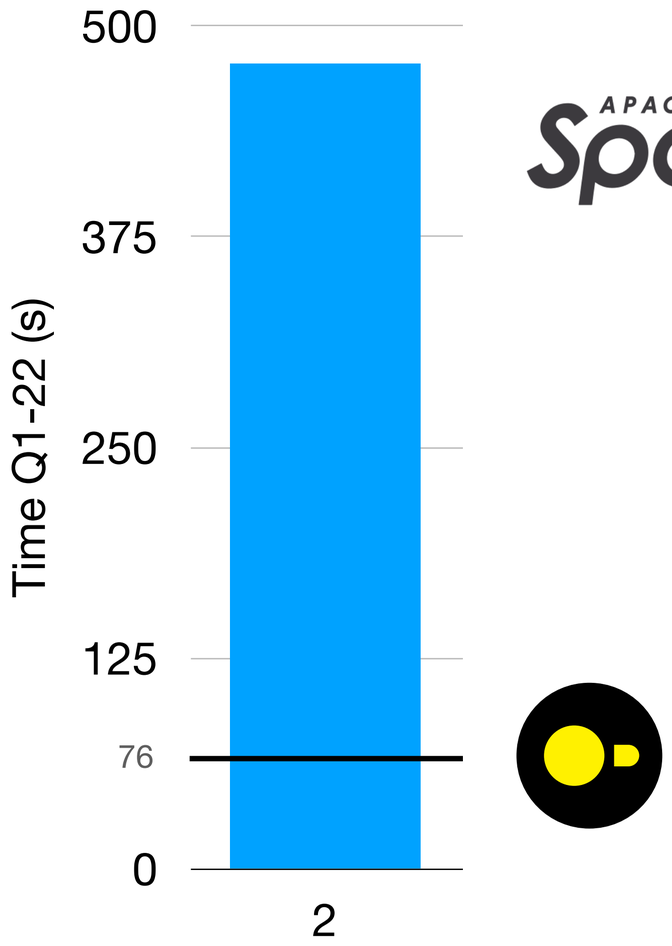
SSDs

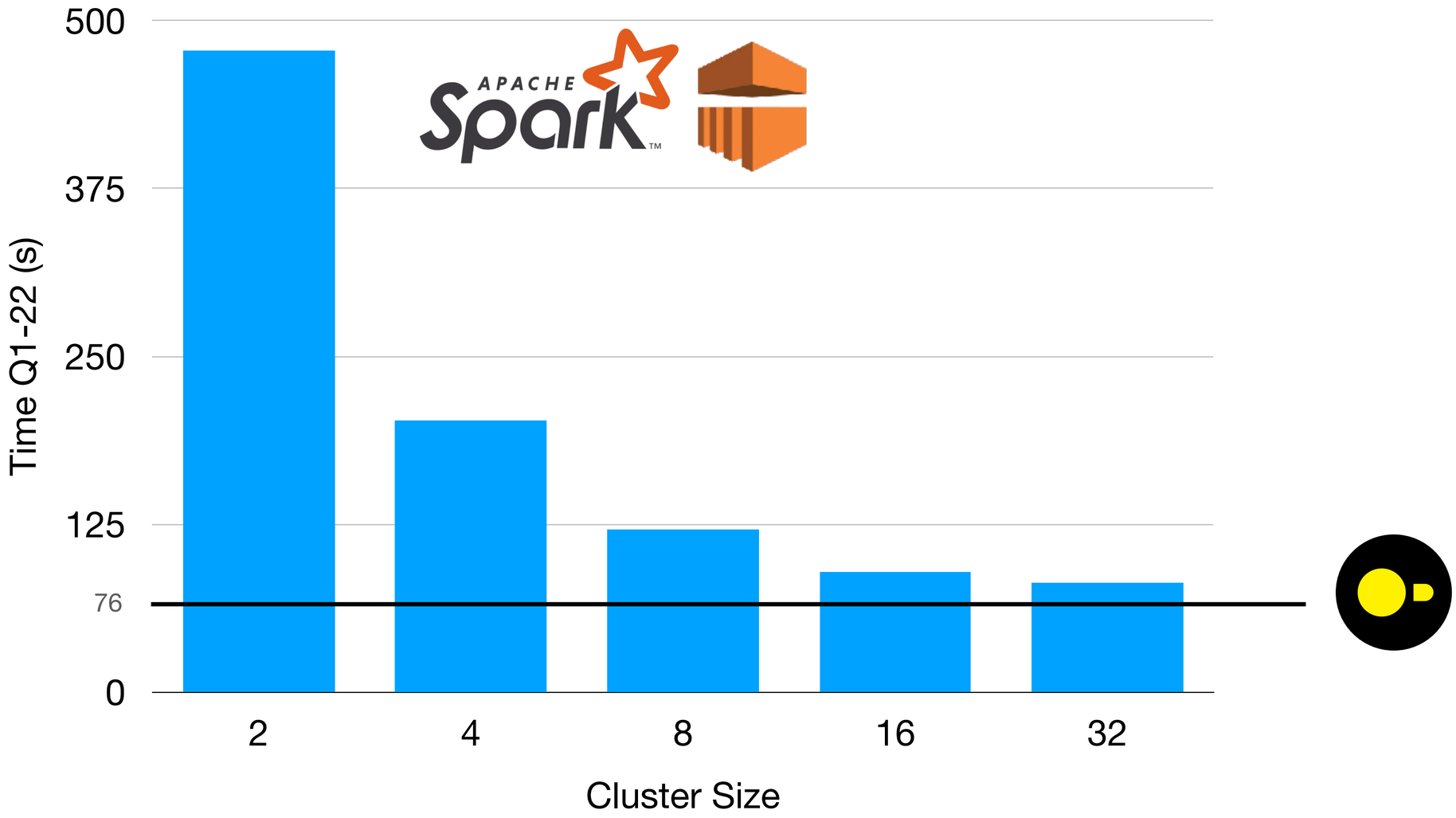
min	max
May 30	May 31

min	max
June 1	June 2

date	id	type	station
May 30			Ams C
May 30			Utrecht
May 31			Ams C
May 31			Schiphol
June 1			Schiphol
June 1			Utrecht
June 1			Utrecht
June 2			Ams C

Experiments!


500
375
250
125
76
0
Time Q1-22 (s)
2


500
375
250
125
76
0
Time Q1-22 (s)
2
4
8
16
32
Cluster Size

A Short Summary of the Last Decades of Data Management

- Tables are eternal
- NoSQL was a bad idea
- Relational systems eat most things
- Big data is dead
- DuckDB

hannes.muehleisen.org
@hfmuehleisen