

Reutlingen
University


13.03. – 17.03.2023, DBKDA 2023, Barcelona

Challenges & Trends in DBMS – Ten Years After

*Fritz Laux, Emeritus Professor at
Reutlingen University*

*Malcolm Crowe, Emeritus Professor at
the University of the West of Scotland*

© F. Laux



Every 5 years, a group of leading DB researchers meet and discuss about Challenges and Trends in DB development.

In this talk we will investigate their reports and compare these research trends with my personal view from 2010.

This talk is not objective, it reflects our very personal view on DB trends of the last 10 years.

My colleague, Malcolm Crowe, is Emeritus Professor at the UWS. He was department head from 1985-1999 and Associate Dean of Faculty 1999-2006,.

Since 2001 he developed Pyrrho, a memory resident DBMS with optimistic concurrency control and true serializability.

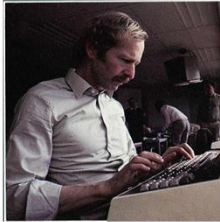
My name is Fritz Laux. I was teaching DB systems and data analytics for nearly 30 years at Reutlingen University.

Together with Malcolm we were involved in DBTech, a European project and initiative of database lecturers, and we published a series of papers and other contributions at IARIA conferences.



Reutlingen
University

Fritz Laux



Once upon a time ...
~ 40 years ago

- *Education: MSc (Diplom) and PhD (Dr. rer. nat.) in Mathematics*
- *Working as SW-analyst, designer and architect of commercial information systems for ZF, Porsche, Bosch/ Junkers, Telekurs, and Swiss PTT*
- *Full Professor for Database and Information Systems at Reutlingen University. Dean of Studies. Supervised >200 Bachelor and Master students, and 3 Ph.D. students*
- *Cofounded DBTechNet (www.dbtechnet.org) and involvement in EU-projects DBTech.pro & DBTech.ext*
- *Research activities in Database Modelling, Transaction Processing, Data Warehousing, and Data Mining.*
- *Research Award, IARIA fellow.*

2 /28

© F. Laux

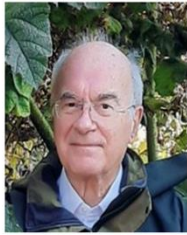


Reutlingen
University


3 /28

© F. Laux

Malcolm Crowe



- *Education: D.Phil. in Mathematics at the University of Oxford in 1979.*
- *Malcolm Crowe is an Emeritus Professor at the University of the West of Scotland, where he worked from 1972 (when it was Paisley College of Technology) until 2018. He was appointed head of the Department of Computing in 1985 -1999 and was appointed Associate Dean of Faculty 1999-2006.*
- *His funded research projects before 2001 were on Programming Languages and Cooperative Work.*
- *Since 2001 he has worked steadily on [Pyrrho DBMS](#) to explore optimistic technologies for relational databases and this work led to involvement in DBTech, and a series of papers and other contributions at IARIA conferences with Fritz Laux, Martti Laiho, and others.*
- *IARIA fellow.*

 <p>Reutlingen University</p> <p>4 /28 © F. Laux</p>	<h3>Motivation</h3>
---	---------------------

↪ *With a history of more than 50 years database research has matured and strongly influenced commercial SW industry*

↪ *DB research and products have extended from core DBMS to multi-purpose systems.*

- ☞ Making DBMS complex and „fat“
- ☞ *Is this useful to add constantly more functionality to a DBMS?*
- ☞ *How about modular and lean systems?*

↪ *Since 1989 ~ every 5 years major DB academics gather and reflect on DB research, trends, and their community's impact*

- ☞ To a great extent they echoed the current hype/trend
- ☞ *How precise has been their self assessment?*
- ☞ *Is it possible to guide and predict future developments?*
- ☞ *Should the reports have given more specific direction for research?*

↪ *More than 10 years ago, I presented my wishes for the DB community*

- ☞ *How do these compare with the reports of the leading DB researchers and what worked out and what not?*

Database technology has strongly influenced and even shaped commercial SW systems.

DBMS have extended from core to multi-purpose systems comprising Web- and application servers, DWH, and OLAP systems.

The DB reports we want to present and analyse gave reason to ask the following 5 questions:

- (1) Is this useful to add constantly more functionality to a DBMS?
- (2) How precise has been the self assessment in the reports and how about their research proposals?
- (3) Is it possible to guide and predict future developments by trend analysis?
- (4) Should the reports have given more specific direction for research?
- (5) How do my wishes from 13 years ago compare with the reports of the leading DB researchers and what worked out and what not?


 Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 5 /28 © F. Laux	Outline
	<p>↳ <i>DB Trends & Research 2005 – 2022</i> ☞ Reports from leading DB researchers</p> <p>↳ <i>Analysis of Trends</i> ☞ Sustained trends or hype?</p> <p>↳ <i>Analysis of Research proposals</i> ☞ What worked out and what not?</p> <p>↳ <i>Challenges for DB researchers</i> ☞ My personal view from 2010</p> <p>↳ <i>Recommendations:</i> ☞ What can we learn from (commercial) trends?</p> <p>↳ <i>Conclusion: quo vadis DB Community?</i></p>

The research reports were authored by 20 – 30 leading DB experts in the last 18 years and compare these research trends with my personal view from 2010.

What were sustainable trends or hypes, what worked out and what not?

What can we learn from commercial trends? Or, should academic research give more direction of the sort I presented already in 2010?

Finally, we give some recommendations to the DB Community from our perspective.


Reutlingen
University

Outline
DB Reports
Analysis
My DB vision
Products
Big Live Data
Comparison
Observation
Conclusion
References

6 /28
© F. Laux

Database Reports from leading experts

↪ *20 to 30 DB experts have assessed DB research and discussed key directions for DBMS development*

- ☞ 2005: The Lowell DB Research Self-Assessment (2003)
- ☞ 2009: The Claremont Report on DB Research (2008)
- ☞ 2016: The Beckman Report on DB Research (2013)
- ☞ 2022: The Seattle Report in DB Research (2018)


↪ *Let's look at the reports from the last 18 years*

- ☞ Distinguish sustained trends from hype
- ☞ Compare research proposals with publications

These are the 4 reports we are going to look at in detail. Between 20 to 30 of eminent DB researchers gathered at certain locations. The reports are named after the reunion place.

During the meeting they identified and discussed database development trends. Their evaluations were compiled into a research and self assessment report. Usually the report appeared 2 years after the meeting (year of meeting in parentheses).

Let us now look at these reports in sequence and assess whether they have identified long term trends or only hypes and compare the research proposals with publication activities.

2005: Lowell DB Research Self-Assessment (Meeting 2003)		
 Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 7 / 28 © F. Laux	↳ <i>Identified trends</i>	<i>Realised trends</i>
	<ul style="list-style-type: none"> ☞ Enormous data sources from Internet apps/e-commerce and science ☞ Micro-sensor technology growth prognosis: 11% /year 	ongoing 2003: 35 bn, 2020: 82 bn actual 5 -6% growth rate/year
	↳ <i>Research proposals</i>	
	<ol style="list-style-type: none"> Integration of text, (sensor) data, information fusion ← motivated by growing data sources <ul style="list-style-type: none"> ⇒ Achieved Result (2005): testbed for XML data only, Not considered: Privacy, Mix of data formats Unsupervised data mining ← motivated by a Fortune 500 survey <ul style="list-style-type: none"> ⇒ The idea was that a background algorithm should mine through the database and find unexpected "pearls of wisdom". ⇒ Achieved Result: There existed already many algorithms, e.g. K-means (MacQueen 1967) ⇒ But: these have never been integrated to create a background process. Privacy ← motivated by growing internet data <ul style="list-style-type: none"> ⇒ Achieved results: the EU law GDPR (2016/18) , UN resolution on the right to privacy (2013), technical proposals: access control, role concept, encryption, obfuscation, etc , ⇒ Addressing multiple data sources, owners, stakeholders → Live Data Integration Self adaptation & repair <ul style="list-style-type: none"> ⇒ Achieved results: Conference proceedings 2013 & 2014 on Adaptive self-tuning systems, but no publication for DBMS, One publication on self healing (2013), but none for DBMS 	


The Lowell meeting took place in May 2003 near Boston, Ma.

It was motivated by a Fortune 500 survey which unveiled the demand for discovery of unexpected "pearls of wisdom" (something new and interesting).

The participants realised that data sources from the internet generate enormous data volumes. The term "Big Data" was not invented yet. Another identified trend was micro-sensor technology where the Fortune 500 survey expected a yearly growth rate of 11%, but up to now only about 5-6% have been realised.

The participants proposed 4 main research areas:

- (1) **Data integration** of all sorts of data to achieve full information fusion. It was motivated by the continuously growing number of data sources. Only one year later a testbed for XML data was developed, but no other formats of data had been considered.
- (2) **Unsupervised data mining** was motivated by the Fortune 500 survey. The vision was just to feed in large amounts of (confusing) data to an algorithm running in the background and expect some valuable insights. Already before 2003 there existed a considerable amount of algorithms for unsupervised data mining (e. g. K-means, DBSCAN, DENCLUE (density based), etc.). To my knowledge, none of these algorithms had been implemented as a background process.
- (3) Better **privacy** was a main concern at the meeting because the internet created a urgent need. The report requested technical efforts for better privacy for all stakeholders when combining data from multiple sources without giving details.
Major non-technical results of this proposal were the EU General Data Protection Regulation (GDPR) and the UN resolution on the right to privacy in the digital age (resolution 68/167).
- (4) **Self adaptation and repair** is a long-standing desire. There have been conferences on these topics. Some DBMS products promise some self tuning and fault tolerance. But none provides self repair or healing.

2009: The Claremont Report on DB Research		
 Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References	↳ <i>Identified trends</i> <ul style="list-style-type: none"> ☞ Enthusiasm for Big Data, extraction of structured & unstructured data ☞ Expanded developer & productivity demands ☞ Architectural shifts in computing (cloud computing, Flash memory, massively parallel computing) 	<i>Realised trends</i> ongoing trend. Data extraction from blogs, e-science, digital social networks and keyword search in Web pages mostly domain specific languages (R) & frameworks (MapReduce 2004/10, PyTorch 2016) many frameworks (.NET, Angular, Spark, Ruby on Rails, LINQ, Spring, Hadoop, etc.)
	↳ <i>Research proposals</i> <ol style="list-style-type: none"> 1. Revisiting database engines ← motivated by DWH/OLAP <ul style="list-style-type: none"> ⇒ Analytics engines (read-mostly, large join and aggregation workloads) ⇒ Designing systems with query optimization for clusters of many-core processors 2. Declarative programming for emerging platforms ← motivated by productivity demands <ul style="list-style-type: none"> ⇒ Developing more powerful and efficient languages to address complex problems ⇒ Need efficient compilers and runtimes, optimize code automatically (JIT compilers) ⇒ Achieved results: functional languages are not satisfying (Datalog, Erlang (1987), Haskell (1990)) 3. Interplay of structured and unstructured data ← motivated by Big Data trend <ul style="list-style-type: none"> ⇒ Integration of all kinds of data over many repositories (managing dataspace) ⇒ Extract structure and meaning from unstructured data over heterogeneous sources 4. Research on Cloud data services ← motivated by Cloud Services <ul style="list-style-type: none"> ⇒ More powerful Cloud APIs, SQL functionality, improved manageability ⇒ Hardware level VM vs multi-tenant hosting 	


When the Claremont Meeting in Berkeley, CA, was held in May 2008 there was some excitement about **Big Data**. This is still an ongoing trend but now (in 2023) the emphasis is on data analysis and AI. A plethora of Data Extraction & Workflow Tools have been developed: IBM Aspera, Astera Unified Data Pipeline, Scrapestorm (AI powered data extraction), Klipa (document processing). A lot of data moving is required for these ETL-tools. There is no coherent solution for all data formats or freshness of information.

To cite the 2009 Claremont Report: “Unlike previous work on information integration, the challenges here are that we cannot assume we have semantic mappings for the data sources and we cannot assume that the domain of the query or the data sources is known. We need to develop algorithms for providing best-efforts services on loosely integrated data. The system should provide meaningful answers to queries with no need for manual integration and improve over time in a pay-as-you-go fashion as semantic relationships are discovered and refined”. This is a nice desire, but if you don’t have the semantics behind it is a big risk to let the computer guess.

The report identified the continuing need for productivity in SW-development, which had been notorious since the 1968 NATO SE-conference in Garmisch. Many new languages and productivity tools had been developed before 2000, so this was not a new trend, and the real problems in program logic remained elusive. New hardware and computing models require **architectural shifts** supporting frameworks like MS .NET (Middleware, Library, Runtime), Apache Spark (cluster computing), Spring (Framework for Java Web Development) and others.

The **research proposals** in this report emphasise (1) **revisiting database engines** due to changed requirements mainly from OLTP to OLAP and to data analytics. Motivated by the demand for productivity the participants proposed (2) **declarative programming** which again was not new nor a surprise.

(3) Research on **managing all kinds of data** over various repositories is motivated by the big data trend and more research on (4) **cloud data services** should lead to more powerful Cloud APIs and improved manageability like hardware level VM and multi-tenant hosting. In my view most research on the **concepts for cloud computing**, e.g. parallel and distributed processing or virtual systems had been accomplished already decades before.

	2016: The Beckman Report on DB Research	
Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 9 / 28 © F. Laux	Identified trends <ul style="list-style-type: none"> Big data trend continued and became democratized rediscovering ACID integration and analysis of data from diverse sources 	Realised trends <p>cheaper storage and processing costs: BigTable, Cassandra, Hadoop/Hive. VoltDB, Neo4J.</p> <p>Scrapestorm, Monarch, Klippa, but ignore provenance, privacy, ownership</p>
	Research proposals <ol style="list-style-type: none"> Scalable big/fast data infrastructure ← motivated by Big Data trend <ul style="list-style-type: none"> Parallel and distributed processing, Query processing and optimization New hardware (GPU, FPGA, ASIC) Diversity in data management ← motivated by Big Data trend <ul style="list-style-type: none"> Different data types, sizes, representations → multiple classes of systems Set oriented parallel processing → declarative query languages Cross platform integration → integration frameworks Cloud services ← motivated by commercial trend <ul style="list-style-type: none"> Elasticity: network latency & bandwidth → can replication help? Transactions & Analytics in the same cloud → data replication? Data sharing: How to support data curation and provenance collectively? 	


The next meeting took place at the Beckman Center, at the Univ. California-Irvine, south of LA in Oct. 2013. The participants quickly converged again on Big Data as a defining challenge due to the availability of cheaper storage and processing costs, and finally because the handling of data had become possible for ordinary users (“democratized”).

The ACID properties had been “rediscovered”. Schema-less data has led to the development of NoSQL systems. There are many such systems providing only weak atomicity and isolation guarantees, making it difficult to build reliable applications and analyze the data. As result, a new class of big data system had emerged that provides full-fledged database-like features. VoltDB is only one example and Neo4J originally a schema-less graph database supports database consistency rudimentarily by constraints (avoids lost updates, but not read anomalies).

Data integration and analysis remained a challenge despite the many data ETL tools because privacy and ownership was not addressed properly.

Motivated by these trends the participants of the meeting made the following research proposals:

- (1) Develop a scalable, fast data infrastructure for big data. The infrastructure should make use of new hardware (GPU, FPGA, ASIC) and leverage well studied concepts of parallel and distributed processing. The proposed research was late. MapReduce (2004) and Apache Spark (2013) had been developed already before the report appeared.
- (2) Support diversity in data management. This proposal aims to address the challenges of cross platform integration and variety of formats. Today, data are often stored in different representations managed by different software systems.
- (3) The report demands research on elasticity, transactions and data sharing for cloud services, which is motivated by commercial trends to move applications into the Cloud. Specifically, the research should investigate if and how data replication can help. Other issues are data curation and provenance when data sharing is desired.

	2022: The Seattle Report on DB Research	
Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 10 /28 © F. Laux	Identified trends <ul style="list-style-type: none"> Big data trend continues with data science (AI and ML) Data governance, ethical & fair use of data Hybrid transactional/ analytical systems (HTAP) 	Realised trends <ul style="list-style-type: none"> TensorFlow, PyTorch, GPT-3, LLM Jupyter, Spark, Zeppelin EU GDPR (2016) / UN DG (2019): Data privacy, ethics & protection Azure/Synapse, Oracle Dual Format, SQL Server/Hekaton
	Research proposals <ul style="list-style-type: none"> Data science ← motivated by data science trend <ul style="list-style-type: none"> Pipeline from raw data to knowledge, Data context and provenance Declarative programming, Metadata management Data governance ← motivated by UN DG, EU GDPR <ul style="list-style-type: none"> Ethical data science, responsible data management → FATE Cloud services ← motivated by commercial trend and HTAP <ul style="list-style-type: none"> Serverless data service (FaaS) with transparent SLA Disaggregated architecture for OLTP and analytic workloads (HTAP) Internet-of-Things (optimisation for distributed data processing) Seamless mix of on-premises (local) and on-demand (Cloud, remote) computing Database engines ← motivated by HTAP <ul style="list-style-type: none"> Heterogeneous computation with GPU, FPGA, SSD → new algorithms Distributed transactions (trade-offs between consistency, isolation, availability, latency) Data lakes (with disaggregated architecture) In-database ML and auto-tuning → the same was proposed 2005 for Data Mining 	

The last meeting of this sort so far took place in Seattle, WA, in Oct. 2018. The report appeared after discussions at SIGMOD and VLDB conferences in August 2022.


For the third time Big Data was identified as a continued trend, now driven by data science using AI and ML. The most spectacular example is GPT-3 (= Generative Pre-trained Transformer) based on a autoregressive language model that uses deep learning to produce human-like text. TensorFlow is a software library for machine learning and artificial intelligence. The PyTorch framework aims to support machine learning applications and Jupyter is used as a scientific data analysis platform. Apache Zeppelin enables Web-based, interactive data science.

The privacy aspect of data governance was mentioned the first time in the 2005 report and resulted already 2016 into the EU GDPR (= General Data Protection Regulation) and recommendations of the UN DG (= Development Group). Now, the focus was on ethical & fair use of data. The report requests to implement metadata annotations and provenance must accompany data items when data is shared.

Motivated by the latest database products like Oracle Dual Format, MS SQL Server/Hekaton and Azure/ Synapse hybrid transactional analytical systems (HTAP) have been identified as an important trend.

This led to 4 research proposals:

- (1) Data Science. Reconsidering the data pipeline from raw data to knowledge, data context and provenance, metadata management and declarative programming. These topics are far from being new. We mentioned it in previous reports already.
- (2) Data Governance. This has not only technical but also social and political dimensions. Governments and enterprises have picked up the topics Fairness, Accountability, Transparency, and Ethics (FATE) often in conjunction with AI.
- (3) Cloud Services. After mastering the core functions of Cloud Services the focus is now on serverless data services, known as FaaS, disaggregated architectures, and IoT. Research is also required on seamless mix of local and remote computing services.
- (4) Database Engines. It is evident that new hardware requires new algorithms. Distributed transactions need to make trade-offs not only between consistency, availability, and partition tolerance (CAP-theorem) but also with isolation and latency. Loosely related data sources (data lakes) and disaggregated processing, in-database ML, and auto-tuning are still far from having matured.

 Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 11 /28 © F. Laux	Overview of Trends & Challenges from the Reports					
	Trend / Motivation	Research proposal	2005	2009	2016	2022
	Data Mining (DM) Technology	unsupervised DM	X			
	Internet privacy	Privacy	X			
	Admin shortage	Self adaptation/repair	X			
	Data integration (any type and format)	Data integration Sensor data	X X	X X	X	
	Productivity demand	Declarative lang. Query optimization Simpler API	X X	X X X	X	X X (ML, SaaS)
	Cloud Computing	Every aspect		X	X	X
	New HW	architecture, algorithms, distributed SW,		X	X	HTAP
	Data Science (DS)	Apply DB concepts to DS				X
	Governance	Unspecified				X

This table summarises the trends and research proposals from the DB reports for a time span of nearly 20 years.


It is remarkable that the 2003/2005 topics, unsupervised DM, privacy, and self adaptation/repair do not reappear in later reports. We will look at possible reasons for it on the next slides.

Data integration was addressed in 2005 and 2009 with special emphasis on sensor data and mixing unstructured with structured data. In the next report 2016 the scope was expanded with the analysis of integrated data from diverse sources.

A consistent demand for better productivity in SW development was shown in all reports with repeated research proposals for high level declarative languages, (automatic) query optimization, simpler and flexible APIs.

Starting with the 2009 report, Cloud computing und new hardware entered the trend scene. New hardware affects architecture, algorithms, and distributed SW, which were identified as research topics.

In my view the latest report from 2022 was a disappointment with the least focus. It reiterated most topics from previous reports under the general labels “Data Governance”, “Cloud Services”, and “Data Science”. For instance data privacy was treated again under “Data Governance”. Auto tuning, cost models, SLA, data analysis/mining using disaggregated computing, IoT, and many others are covered under the label “Cloud computing”. A complete loss of depth and focus.

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>12 /28</p> <p>© F. Laux</p>	<p style="text-align: center; background-color: yellow;">Comment</p> <p>↪ <i>The DB Reports identified commercial or technological developments and oriented their research proposals in line with these trends</i></p> <ul style="list-style-type: none"> ☞ This makes some sense in order to give theoretical and conceptual support for these trends ☞ The disadvantage of this approach is that long-standing problems are not fundamentally addressed ☞ Risk of short-lived hype themes <p>↪ <i>The alternative would be to identify such long-standing problems (so called “Grand Challenges”) and propose research for a solution.</i></p> <ul style="list-style-type: none"> ☞ As example for such an approach we recall later in this talk a keynote from DBKDA 2010 ☞ It is intended to provide a contrast to the DB Reports
---	---


The database reports mainly identified and reflected commercial or technological developments. The research proposals were in line with these trends. This makes some sense in order to give theoretical and conceptual support for these trends. The disadvantage, however, of this approach is that long-standing problems are not appropriately and fundamentally addressed. There is a risk that research is proposed on short-lived hype themes. We will discuss these issues on the next slide.

Some of the trends like “Big Data”, “Data Science” or “Cloud Computing” had enough momentum to drive research without academic impulse.

The alternative would be to identify such long-standing problems, so called “Grand Challenges”, and propose research on these. “Grand Challenges” are difficult but important problems that are more than ordinary research questions or priorities. They are characterized by

- (1) probably solvable within 10~15 years
- (2) broad application field with social and economic relevance, and
- (3) generally require interdisciplinary knowledge.


Solving a “Grand Challenge” would bring important progress to society. Its solutions require fundamental new ideas, like the relational data model was for database systems.

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>13 /28</p> <p>© F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Analysis of long term Trends vs short Hypes</h2> <p>↳ <i>Long term trends/challenges</i></p> <ul style="list-style-type: none"> ☞ The analysis of large data sets originates from 1970s with new names every 10 years or so: <ul style="list-style-type: none"> ⇒ MIS (1970+) – DSS (1980+) – OLAP (1990+) ...– Big Data (2010+) – Lakehouse (2020+) ☞ Over this period the DBMS community was occupied with data integration and analysis ☞ There are many products aiming to solve the problem. They rely on ad hoc solutions (drivers, connectors) and predefined analysis. No coherent concept is offered. ☞ The DB reports give no clue how to solve the integration problem <p>↳ <i>Hypes</i></p> <ul style="list-style-type: none"> ☞ Focus shifts to other trends or previous research: e.g. unsupervised DM, Privacy, self adaptation/repair ☞ Some trends vanish and reappear under new names after some time like: Expert System → Machine Learning, Self Driving Car → Autonomous Vehicles ☞ Some have been disgraced and vanished: Knowledge Management, Virtual Reality (e.g. Second Life) , or reappear?
---	---

The analysis of large data sets from different data sources is a long-standing problem that originates from 1970s with new names every 10 years or so. The first name was probably Management Information System (MIS). Then we had Decision Support Systems (DSS), Data Warehouse and OLAP Systems, then Enterprise Information Systems (EIS) and Big Data. The names and the focus of analysis changed over the decades, but the integration problem remained with all its facets. The latest name for this challenge seems to be Lakehouse, a loosely related set of data with different formats and from various sources.

The DBMS community was focusing on data integration with changing intensity and many products have been developed to help with this task. But no coherent framework was available so far. The DB reports addressed data integration multiple times but gave no hint how to solve the problem.

Often it is not easy to identify a serious, lasting trend and distinguish it from hypes that vanish soon again. The possible implications of a new idea can give guidance what will remain and what will disappear. But sometimes the idea was too early and technology was not yet ready for it. E. g. MIS (1970s), AI/Expert systems (1980s). Gartner gives more examples for Hype Cycles. But in any case for a successful trend its (research) challenges need to be solved.

 Reutlingen University Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References 14 /28 © F. Laux	Publications on Topics from the DB Reports						
	Search term	DBLP	IEEE	ACM DL	Think Mind	Research	Since year
	unsupervised data mining	62	45	62	48	unsupervised Data Mining	2005
	Internet privacy	1532	272	145	370	Privacy	2005
	Self adaptation & repair	0	1	19	1	Self adaptation/ repair	2005
	Data integration struct. unstruct.	9	74	910	7	Data Integr. Sensor data	2005 2009
	Productivity demand	5	2	5	0	Declarative lang. Query optim. Simpler API	2005 2005 2009
	Cloud comput.	13466	79007	18400	397	Every aspect	2009
	HTAP	33	29	135	0	architecture, algorithms, distributed SW,	2022 2009 2016
	Data Science (DS)	6354	16905	8461	1980	Apply DB concepts to DS	2022
	D. Governance	421	268	502	89	Unspecified	2022

Analysing the reports in the light of publications that picked up the research proposals we must say that publications of most topics started well before the report appeared. This is partially because the reports were published approximately 2 years after the meetings and the participants are academics who publish what they think is an important trend or they already research on.

The publication count was the result of a search with all keywords AND –connected or as a string of adjacent terms. For example “data mining” was used quoted instead of unquoted. The only exception was DBLP because it did not support string search, only word search. The search always started with the publication year of the DB report.

There was research for unsupervised DM, but not running as a background service. The Lowell report of 2005 did not lead to a commercial implementation and I could not find papers addressing this research proposal. Maybe there was no need for it or processing cost are too high compared to the results.

Similarly self adaptation and repair lacks a general solution, only partial resp. ad hoc solutions have been built into SW-systems.

There is much ongoing research on internet privacy. The major part of the publications addressed technical measures for privacy, while some publications also addressed social and political dimensions of privacy. In my view the technical issues are solved even if measures are complex. The main problem are individual behaviour and political regulations. The 2005 report was late to request research on internet privacy. About 20% of all publications on internet privacy appeared already before 2005.

Data Integration of structured and unstructured data picked up interest mainly after 2005, since before that time data integration was focused on the integration of structured data. It is interesting to note that there exists no general (commercial) framework for data integration providing the already mentioned data economy, provenance, and privacy.


The long-standing demand for better productivity of SW development did not result in a lot of papers. I feel that the call for research was justified and necessary. It is unclear why not more effort is put into a higher productivity, instead the number of languages, APIs and SW-frameworks grow and become more and more complex.

Cloud computing is a megatrend since approximately 2010 and still produces a lot of papers. Yet, most of the concepts needed, e.g. distributed architecture, parallel processing, data transfer, and synchronisation have been investigated many years ago and seem to be mostly solved.

A rather new trend “hybrid transactional and analytical processing” (HTAP) has become an important business segment.

Data Science and Governance have been identified as new megatrends in the latest report. The latter is mainly a question of legal regulation, which clearly needs support from IT with user authorisation, access control, and activity logging producing more sensitive data.

Data Science with exploitation and analysis of big data by means of DB concepts, ML and AI methods picks up momentum again with large AI models

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>15 / 28</p> <p>© F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Publication Analysis of proposed Research (1/2)</h2> <ul style="list-style-type: none"> ↳ <i>Privacy , Cloud Computing, and Data Science are by far the most published topics</i> ↳ <i>Privacy – 2319 publications from 2005 to 2022</i> <ul style="list-style-type: none"> ☞ seems to be an ongoing research task ☞ Technical aspects of Privacy/Encryption are largely studied ↳ <i>Cloud Computing – 111270 publications from 2009 to 2022</i> <ul style="list-style-type: none"> ☞ Reflects the business impact rather than conceptual advances ☞ All major concepts are well investigated ↳ <i>Data Science – 33700 publications in 2022</i> <ul style="list-style-type: none"> ☞ The new hype theme with high potential and impact ☞ Before the meeting a first architecture was already published by OpenAI in May 2018, the idea is much older ☞ GPT-3 and HyperCLOVA demonstrated impressive results ☞ Examples <ul style="list-style-type: none"> ⇒ generate query-specific code for SQL processing (CodexDB, GPT-3) ⇒ Article writing for newspaper Guardian (GPT-3), marketing slogans and tourist guidance (HyperCLOVA) ⇒ Screening for early signs of Alzheimer's disease
--	---

Privacy , Cloud Computing, and Data Science are by far the most published topics.

There have been 2319 publications on Privacy from 2005 to 2022. From its reoccurrence in the reports we may conclude that this issue is of continuing interest. However, the papers mostly focus on technical aspects, whereas with Data Governance the real need is for legal regulation and skilled users.

Cloud Computing is a long-lasting hype and business which created 111270 publications from 2009 to 2022. It reflects the business impact rather than conceptual advances. All major concepts are well investigated.


Data Science emerged as a new trend and research topic with 33700 publications in 2022. The new hype theme has high potential and impact. It is fueled by new and cheaper hardware, increased processing power and distributed algorithms for parallel processing.

Before the meeting a first architecture was already published by OpenAI in May 2018, the idea is much older. GPT-3 and HyperCLOVA demonstrated impressive results for generating understandable text.

Examples:

generate query-specific code for SQL processing (CodexDB, GPT-3)


Article writing for newspaper Guardian (GPT-3), marketing slogans and tourist guidance (HyperCLOVA), Screening for early signs of Alzheimer's disease.

 <p>Reutlingen University</p> <p>Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References</p> <p>16 /28 © F. Laux</p>	<h3>Publication Analysis of proposed Research (2/2)</h3> <ul style="list-style-type: none">↳ <i>Every report since 2005 requested research for better productivity, resp. declarative languages</i><ul style="list-style-type: none">☞ New languages and frameworks have been developed☞ Productivity hasn't advanced much, but complexity has!↳ <i>The DB reports were mainly driven by IT trends and less based on conceptual research needs</i>↳ <i>Take the Cloud Computing trend as example</i><ul style="list-style-type: none">☞ The fundamental theory (distributed and parallel processing, networking protocols, synchronisation, etc.) has been developed long ago.☞ What is needed is research on architectural concepts for different services, isolation, fault-tolerance, performance (SLA), and privacy.
---	--

Every report since 2005 requested research for better productivity and declarative languages. New languages and frameworks have been developed and are in use. But productivity hasn't advanced much because complexity has increased! And after all the SW lifecycle outlives technology which consumes a lot of human resources to maintain old SW.

The DB reports were mainly driven by IT trends and less based on conceptual research needs.

What is still needed is research on architectural concepts for different services, isolation, fault-tolerance, performance (SLA), and privacy.

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>17 / 28</p> <p>© F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Recommendation: What the Research should Address</h2> <p>↳ <i>Research proposals</i></p> <ul style="list-style-type: none"> ☞ The research proposals were often late, but some trends have resulted in impressive business success <ul style="list-style-type: none"> ⇒ The influence of the reports remain unclear ⇒ No hint was given in the DB reports how to tackle the topics <p>↳ <i>What would have helped?</i></p> <ul style="list-style-type: none"> ☞ It could have been expected that a research community would give a more detailed analysis and propose directions to a solution <ul style="list-style-type: none"> ⇒ The early reports up to 1995 have done better ☞ A stronger analytical view on the trends would have been necessary in order to give guidelines for conceptual research <ul style="list-style-type: none"> ⇒ to justify the trends or ⇒ warn of pitfalls
--	--

Topics were picked up when they had already a trendy momentum. The reports resulted mostly in publications that deal with specific technology.

What is really needed? Not only research on technology!


Research on concepts, models, and paradigms would be necessary to be able to understand and work with new technology that might arise.

A kind of critical view on the trends would have been necessary in order to give guidelines for conceptual research to establish and justify these trends or warn before shortcomings.

It could have been expected that a research community would give a more detailed analysis and propose directions to a solution.

Some important themes were identified as research topics, but no hint was given in the DB reports how to approach the problems.

A stronger analytical view on the trends would have been necessary to distinguish between short hypes and important, long-standing challenges. These challenges must be analyzed and guidance should be offered to justify conceptual research to support a trend with theory, leaving technology to the developers.

 <p>Reutlingen University</p> <p>Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References</p> <p>18 / 28 © F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">2010: My personal research challenges</h2> <p>📌 Presented at DBKDA 2010 : “I have a Dream” – a Vision on Database Technology [Laux2010]</p> <p>📌 The “Grand Challenges” were: <i>to overcome the following limitations</i></p> <ul style="list-style-type: none"> ☞ Schema first requirement → [Laux2010], slides 4 - 9 ⇒ not addressed in the DB reports ☞ Data integration problem → [Laux2010], slides 10 - 17 <ul style="list-style-type: none"> ⇒ repeatedly identified and discussed in the report 2005, also addressed in 2009, 2016, and 2022 reports, but no research direction proposed ⇒ My proposal: Virtual integration preserving ownership and provenance ⇒ The Big Live Data (Crowe/Laux 2017) concept is addressing these issues ☞ Data access/retrieval is for experts only → [Laux2010], slides 18 – 27 <ul style="list-style-type: none"> ⇒ Issue in report 2005 (and mentioned in 2016, 2022 reports with focus on optimization), best effort queries proposed (2009 report), no hints how to simplify access or retrieval ⇒ My proposal : Exact declarative query where possible (schema available) and intentional search using context, ontology, etc. (no schema available)
--	---


In 2010 I gave a talk at DBKDA210 in Les Menuires, France, where I presented my vision about DB developments in the next decade.

The talk concentrated on 3 “Grand Challenges” and gave research directions how to solve the problems.

The „Challenges“ to tackle were threefold:

- (1) Usually a database requires a schema upfront to guarantee consistent and reliable data. Schema-less databases cannot guarantee data quality. Unfortunately, this requirement was never addressed in the reports.
- (2) The necessity to integrate data from different sources and formats was repeatedly discussed in academic papers with different focus. All DB reports since 2005 identified the Data Integration as a research challenge but none gave a hint how to approach the problem.
In my talk from 2010 I was proposing virtual data integration preserving ownership and provenance. In the meantime “Big Live Data” [Crowe2017] is addressing these issues and the Pyrrho DBMS [Crowe2018] is a prototype example demonstrating that the idea is practical and working.
- (3) The 2005 report already identified the need to have an easy to use interface for data access and retrieval. Best effort queries have been proposed in 2009 but these are no answer to many commercial information systems: Banks need exact data as companies need exact data for tax purpose.
My proposal was using a declarative query language when a schema is available and to use intentional search with additional meta-information when a schema is not available.

Let us now look **how** the Data integration problem was addressed in my talk from 2010.



Reutlingen
University

Outline
DB Reports
Analysis
My DB vision
Products
Big Live Data
Comparison
Observation
Conclusion
References

19 /28
© F. Laux

The Data Integration Problem (excerpt [Laux2010], slide 12)

↪ Situation in 2010

- ☞ Analysis of highly scattered but interrelated data requires copying into a single schema (periodical ETL process)
 - ⇒ Data is always out of date
 - ⇒ Much of the data is never used
- ☞ Data provenance and ownership is lost
 - ⇒ Control over data is lost
 - ⇒ Semantic data quality unknown


↪ *How can we perform the required analysis while leaving the data where it is?*

In the 2010 talk the situation with the Extract-Transform-Load (ETL) process was described. At that time data was normally copied from highly scattered enterprise sources into a Data Warehouse (DWH).

This kind of data integration is resource intensive and results in outdated information most of the time. Much of the data is not used by later analyses.

In addition, the ETL process does not care about provenance and ownership. The semantic data quality is unknown and data is only syntactically controlled by the ETL process.

The question in 2010 was: *How can we perform the required analysis while leaving the data where it is?*

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>20 / 28</p> <p>© F. Laux</p>	<p style="background-color: yellow; text-align: center;">The Data Integration Vision (excerpt [Laux2010], slides 13, 15)</p> <p>↳ <i>Database is virtual, but manages an interrelation schema or at its best an integration schema</i></p> <p>↳ <i>In situ Storage</i></p> <ul style="list-style-type: none"> ☞ Data remains where it is created ☞ Ownership and provenance is preserved <p>↳ <i>Data may be cached for performance</i></p> <ul style="list-style-type: none"> ☞ Trade-off between performance and consistency, resp. data freshness <p>↳ <i>Some ideas for storage:</i></p> <ul style="list-style-type: none"> ☞ Data stays at its source location <ul style="list-style-type: none"> ⇒ to preserve provenance ☞ Select as early as possible <ul style="list-style-type: none"> ⇒ Requires query decomposition ☞ Move data only on request <ul style="list-style-type: none"> ⇒ performance boost with caching or hoarding
--	---


The proposal from 2010 was not to move the data to a central DWH but leave the data where it resides and maintain an integration schema in order to know how its semantics and its interrelationship. This allows to combine only data when it is needed for analyses.

This virtual integration was not really a new idea, but with additional information and technical support ownership and provenance could be preserved.

Data could be cached for performance reasons. But this is a trade-off between performance, consistency and data freshness.

Freshness of data need is needed and can be ensured with validation tags and/or modification bits.


To reduce data traffic early query decomposition and caching is helpful.

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>21 /28</p> <p>© F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">The Product Reality</h2> <ul style="list-style-type: none"> ↳ <i>Since 2000 many products have been developed, but mostly ETL or ad hoc solutions (using drivers, connectors)</i> ↳ <i>Examples: (Gartner lists more than 100 products)</i> <ul style="list-style-type: none"> ☞ Informatica PowerCenter (ETL) ☞ SQL Server IS, Oracle GoldenGate, IBM DataStage (ETL and virtual integration) ☞ Denodo (virtual integration) ↳ <i>These products do not address data ownership, provenance, privacy, and data freshness</i> ↳ <i>Malcolm Crowe picked up the idea for virtual data integration</i> <ul style="list-style-type: none"> ☞ Always using the latest data ☞ Preserving ownership, provenance, and privacy ↳ <i>Big Live Data (Crowe/Laux 2017) is an implementation of this idea</i>
---	--

Many products developed since the year 2000 offer ETL or ad-hoc solutions for data integration using various drivers and connectors. Gartner lists more than 100 products from leading database and data warehouse vendors. Some even offer virtual data integration, but none addresses data ownership, provenance, and privacy. Data freshness is not guaranteed per-se if a virtual DWH is used. There is no control or indicator if sources are not up to date.

The best conceptual, rigorous, and most practical answer to data integration seems to be the work of Malcolm Crowe which we call “Big Live Data”. The implementation covers virtual data integration, provides transactional write back possibility (if the owners allow write access). It respects ownership, provenance, privacy, and network traffic is kept to a minimum.

Tests with his prototype demonstrated that the implementation performs well.

 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>22 /28</p> <p>© F. Laux</p>	<h2 style="text-align: center; background-color: yellow;">Big Live Data</h2> <p>↳ <i>The Big Live Data (Crowe/Laux 2017) implementation leverages the following concepts</i></p> <ul style="list-style-type: none"> ☞ The concept of VIEW is suitable - real data held elsewhere ☞ Use REST services –associated with elementary database operations like CRUD ☞ Caching indicators - ETags [Fielding and Reschke, RFC 7232] <ul style="list-style-type: none"> ⇒ “All web servers should provide an ETag string for any request” ⇒ Subsequent requests can supply an ETag as change indicator ☞ A simple implementation of an ETag is the Row-version validation (RVV) [Laiho and Laux, 2010]. It applies to relational data at the tuple level <ul style="list-style-type: none"> ⇒ RVV is an integer that is incremented when any value of a row changes ⇒ Used in DBMS to support optimistic transaction control <p>↳ <i>The contribution is to bring these mechanisms together for the first time</i></p> <ul style="list-style-type: none"> ☞ Building a versioned form of REST Views ☞ Providing a data integrator with relational tools
---	--

Let’s have a closer look on the implementation idea. It leverages the following concepts:

It uses VIEWS to ensure that only cleared data is used and transferred if necessary - real data are held in tables elsewhere.

It uses REST services, which are always associated with elementary database operations like CRUD.

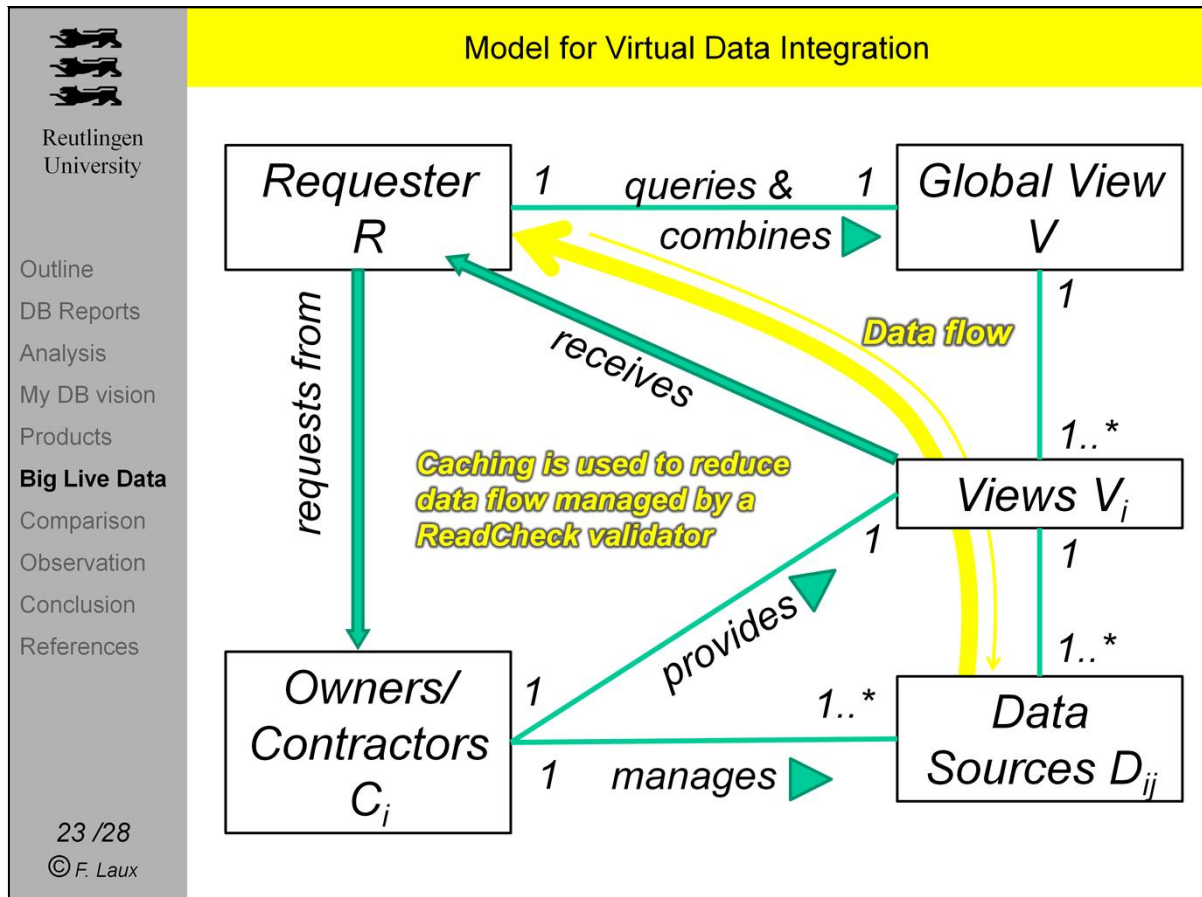
ETags are used as caching indicators, which have been proposed by Fielding and Reschke [RFC 7232] and implemented in the HTTP 1.1 protocol.

“All web servers should provide an ETag string for any request”. A subsequent request can then supply an ETag as a validator for the data.

Row-version validation (RVV, Laiho and Laux, 2010) is an example how an ETag can be implemented for relational data at the tuple level.

RVV is an integer that is incremented when any value of a row changes. It is used in DBMS to support optimistic transaction control and can be used as ETag as well.

The “Big Live Data” idea brings these concepts together for the first time.



This graphical model illustrates how the flow of data requests are processed and how the data flow from the sources to the requester.

The integration is based on the Global-as-View mediator where the global schema is populated by views from the local sources. The requester queries the Global View by sending decomposed queries to the contributors who own the data sources. The owners provide views according to their privacy policy.

The partial result data provided by the contributors are assembled to form a global answer to the request.


This assembly is possible because the requester has access to the global view and therefore knows how to assemble the partial data.

ReadCheck is a validator for data freshness and provides an efficient way of ensuring correct data as of now. Only the data that is needed will be selected from the local views and only the results are transferred and combined at the requester's site.

Similar or repeated request can use the ReadCheck validator to avoid unnecessary data flow if the source view has not changed since the last query.

This mechanism can even be used in long transactions to support ACID properties.

The views on the data sources can be augmented to cover provenance, ownership, and processing restrictions.

 <p>Reutlingen University</p> <ul style="list-style-type: none"> Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References <p>24 /28 © F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Comparison of DB Reports with “Big Live Data” (1/2)</h2> <p>↳ <i>The DB reports are motivated by trends</i></p> <ul style="list-style-type: none"> ☞ The DB proposals are determined by trends ☞ No assessment of the trends in the reports, e.g. no critique of “fat software” ☞ Proposals do not give any hint or guidance how to tackle the problems, some where late <p>↳ <i>Our DB vision is motivated by needs</i></p> <ul style="list-style-type: none"> ☞ Driven by an analysis of problems in IT, long-standing problems ☞ Prefers concepts and abstraction in favor of ad-hoc solution and presents an idea for a solution ☞ The waiver of the “schema first” requirement became “no schema”, not what I wanted! <p>⇒ <i>Our answer to “no schema”: The Typed Graph Model [Laux2020a] and Graph Data Models and Relational Database Technology [Crowe2023]</i></p>
---	--

The DB reports are motivated by trends and thus the proposals are determined by trends.

Trends seem to prefer complex systems over lean solutions. This is the result of commercial products, which aim to always add functionality.

The research proposals do not give any hint or guidance how to tackle the problems, some where late and products ahead.


The reports did not assess the trends, e.g. no critique of “fat software”

Our DB vision is motivated by needs. It is driven by an analysis of problems in IT, mostly long-standing problems.

We favor concepts and abstraction over of ad-hoc solution and present an idea for a solution.

Some ideas have resulted in even more relaxed proposals: These suggested to abandon the schema completely and most graph database products picked it up. This was not what I wanted! We believe that a schema is essential for high quality data.

The Typed Graph Model (TGM) [Laux2020a] shows how a schema can enrich semantics of a property graph model. Malcolm Crowe [Crowe2023] will present an implementation of the TGM based on Pyrrho DB in this DBKDA conference.


 <p>Reutlingen University</p> <p>Outline</p> <p>DB Reports</p> <p>Analysis</p> <p>My DB vision</p> <p>Products</p> <p>Big Live Data</p> <p>Comparison</p> <p>Observation</p> <p>Conclusion</p> <p>References</p> <p>25 /28</p> <p>© F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Comparison of DB Reports with “Big Live Data“ (2/2)</h2> <h3>↪ Comparison</h3> <ul style="list-style-type: none"> ☞ The DB reports received much more attention than our vision! ☞ Nearly every DB report covered “data integration”, but none mentioned “virtual data integration” ☞ Nevertheless, “virtual data integration” had 117 publications in DBLP/ACM/IEEE/ThinkMind, but no coherent solution was implemented ☞ My proposal from 2010 finally led to the “Big Live Data” implementation in 2017 <h3>↪ Both approaches have their merits and drawbacks:</h3> <ul style="list-style-type: none"> ☞ DB reports <ul style="list-style-type: none"> ⇒ Broader scope and attention, more short hypes, considers trends, less guidance, superficial ☞ My vision <ul style="list-style-type: none"> ⇒ Narrow scope on DBMS, mostly conceptual view, less attention, follows “grand challenges” not trends.
---	---

When we compare the DB reports with our “Big Live Data” vision it is clear that the DB reports received much more attention than our vision. This is evident, because 20-30 leading DB academic produced the report, which appeared in the CACM. Nearly every DB report covered “data integration”, but none mentioned “virtual data integration”. Nevertheless, the search term “virtual data integration” had a total of 117 publications in DBLP/ACM/IEEE/ThinkMind, but no coherent solution was implemented. About 1/3 of the publications appeared already before 2010.

The proposal from 2010 finally led to the “Big Live Data” implementation in 2017, which provides a general framework to approach data integration with current data and respecting the important aspects of privacy and provenance.

The DB reports provide a broader scope and attention, but face more short hypes. They consider trends and give less guidance. Some analyses were rather superficial.

My vision from 2010 was narrow in scope and addressed only 3 main themes. It was mostly a conceptual view and followed “grand challenges” not trends. The attention range was small because it was a talk only, not a research paper.



Reutlingen
University

- Outline
- DB Reports
- Analysis
- My DB vision
- Products
- Big Live Data
- Comparison
- Observation**
- Conclusion
- References

26 /28
© F. Laux

Some observations about the “research flood”

↳ *If one searches publication on the proposed research topics the following observations can be made*

- ☞ Minor achievements
 - ⇒ Due to publication pressure
- ☞ Restricted applicability
 - ⇒ Only for a certain technology or application
- ☞ Increase of complexity
 - ⇒ An overly complex environment results in faulty and expensive application
- ☞ What is needed?
 - ⇒ Not only research on technology
 - ⇒ Research on concepts, models, and paradigms, to be able to understand and work with new technology that might arise.


When doing the literature search for the proposed research topics I made some observations that made me question the common publication practice.

I found many papers offering only minor achievements. Others, with preliminary results appeared like a serialized novel. Papers with very narrow focus and only applicable for a specific technology or use case.

This all results in a „flood“ of publications.

From time to time it seems that results fall into oblivion and years later new publications appear applying the forgotten concept in a new setting.

What is really needed is less but substantial publications on concepts, models, and paradigms, to be able to understand new technology that might arise and apply it reasonably.

 <p>Reutlingen University</p> <ul style="list-style-type: none"> Outline DB Reports Analysis My DB vision Products Big Live Data Comparison Observation Conclusion References <p>27 / 28 © F. Laux</p>	<h2 style="background-color: yellow; text-align: center;">Conclusion</h2> <p>↳ <i>We need more concentration on substantial research for</i></p> <ul style="list-style-type: none"> ↳ Lean software systems and how they are developed ↳ modular SW systems ↳ high level languages, simple like keyword search ↳ Tools to check and improve SW quality ↳ Trustworthy systems and explainable AI (XAI) <p>↳ <i>And, we should not forget previous achievements and “reinvent the wheel” again, e.g.</i></p> <ul style="list-style-type: none"> ↳ Data structures ↳ DB models ↳ set-oriented programming and processing
--	---

We need more concentration on real, substantial research on concepts rather than technological specialities.

Lean and modular SW systems would help to manage the whole SW cycle in a better way.

High level, domain specific languages could boost SW development. Tools to check and evaluate SW might improve its quality.

Trustworthy systems and explainable AI are necessary for a broad acceptance of these “intelligent” systems.

Finally, we should not forget previous achievements and “reinvent the wheel” again, e. g. we learned a lot about data structures, DB models, set-oriented programming and processing. We should use this knowledge properly.

Here is only one example:


Set-oriented programming is much more powerful than procedural programming as was demonstrated with Smalltalk convincingly.

The following code example finds all vowels in a text. The code is simple and readable because of a powerful standard class library with collection classes.

Setup: | text | text := 'Mein ganz kleiner Text.' asUppercase.

Code: text asBag select: [:char | char isVowel].

Result: Bag('E(*4)' 'I(*2)' 'A(*1)')

	References
Reutlingen University	[Lowell2005] S. Abiteboul et al., "The Lowell Database Research Self-Assessment", CACM 2005, vol. 48, No. 5.
Outline	[Claremont2009] R. Agrawal et al., "The Claremont Report on Database Research", CACM 2009, Vol 52, No 6.
DB Reports	[Beckman2016] D. Abadi et al., "The Beckman Report on Database Research", CACM 2016, Vol 59, No 2.
Analysis	[Seattle2022] D. Abadi et al., "The Seattle Report on Database Research", CACM 2022, Vol 65, No 8.
My DB vision	[Laux2010] F. Laux, "I have a Dream – A Vision on Database Technology", Keynote at DBKDA 2010, https://www.iaia.org/conferences2010/filesDBKDA10/DBKDA_2010_Speech_I_have_a_Dream.pdf
Products	[Crowe2017] M. Crowe et al., „Data Validation for Big Live Data“, DBKDA 2017, Barcelona, Spain, ISBN13: 978-1-61208-558-6
Big Live Data	[Crowe2018] M. Crowe, F. Laux, "DBMS Support for Big Live Data", Presentation at InfoSys 2018, Nice, https://www.iaia.org/conferences2018/filesDBKDA18/MalcolmCrowe_DBMS_Support.pdf
Comparison	[Laux2020a] F. Laux, "The Typed Graph Model", DBKDA 2020, Lisbon, Portugal, ISBN: 978-1-61208-790-0
Observation	[Laux2020b] F. Laux, "Live Data Integration" Presentation at InfoSys 2020, Lisbon https://www.iaia.org/conferences2020/filesDBKDA20/FritzLaux_Keynote_LiveDataIntegrationSlides_WithNotes.pdf
Conclusion	[Crowe2023] M. Crowe, F. Laux, "Graph Data Models and Relational Database Technology", DBKDA 2023
References 28 /28 © F. Laux	

All the DB reports have been published in the Communications of the ACM (CACM).

In [Crowe2017] the term "Big Live Data" was introduced and concepts for its technical implementation were presented.

The presentation [Crowe2018] illustrates how a DBMS can support Big Live Data with Pyrrho.

The paper of [Laux2020a] introduced the Typed Graph Model (TGM) as an extension of the Property Graph Model (PGM) that was used in the presentation [Laux2020b] as a tool to facilitate the integration of different data structures and models.