# A 20/20 VISION OF THE VLDB-2020?

S. M. Deen [Moderator]
University of Keele, England
deen@cs.keele.ac.uk

Anant Jhingran
IBM Almaden, USA
anant@us.ibm.com

Sham Navathe
Georgia Inst. Tech, USA
sham@cc.gatech.edu

Erich J Neuhold
GMD, Germany
erich.neuhold@gmd.de

Gio Wiederhold
Stanford University, USA
gio@cs.stanford.edu

## 1. Introduction : Moderator

A 20/20 vision in ophthalmology implies a perfect view of things that are in front of you. The term is also used to mean a perfect sight of the things to come. Here we focus on a speculative vision of the VLDB in the year 2020. This panel is the follow-up of the one I organised (with S. Navathe) at the Kyoto VLDB in 1986, with the title: "Anyone for a VLDB in the Year 2000?". In that panel, the members discussed the major advances made in the database area and conjectured on its future, following a concern of many researchers that the database area was running out of interesting research topics and therefore it might disappear into other research topics, such as software engineering, operating systems and distributed systems. That did not happen.

However, in the last 15 years the database research has not been unquestioned. Some concern about its future research is expressed in the Asilomar report [Sigmod Record, Vol (27:4), 1998, p74], as well as by some eminent researchers who fear that the excitement has gone. We must take these criticisms seriously. As we are in the year 2000, this is the right time to look again at database research and to ponder on its future growth over the next 20 years.

To date the database technology has provided large sharable persistent reliable and quality controlled storage and management of data. For many years the focus of the DB research in the VLDB series was dedicated to improving this core technology, which is recently being extended to include the infrastructure for information system development. However, over the last 30 years, the relational model has dominated this technology. By 2020, the 50th anniversary year of that model, we are likely to see many new rich approaches and middleware brought about by the challenge of the Internet/Web, new application domains and new hardware technology.

Additionally, storage and handling in the next 20 years will include data, processes, execution strategies quality of service, relevance and ease of search, in addition to the traditional concerns of reliability and optimisation. There will be a growing need to interact with thousands of databases in the Internet/Web, requiring a database capability to cooperate and negotiate to reach compromises based on preferences. Associated with this, there will be an enhanced need for security, semantics and interoperability, as already being encountered in many distributed applications. There is also likely to be an impact of newer technologies, such as molecular computing and quantum computing.

The 1986 panel identified relational model as the only major achievement in data modelling, without any parallel. SQL was viewed as a great success, eclipsing many other offerings of that time. Many research trends current at that time were predicted to continue as rewarding (rather than dry) topics of investigation (as they actually did), but the expectation of great advancement in information (including data, images, sound) bases, "intelligent" and natural-language processing and image queries have not been materialised. Furthermore, the enthusiastic (encouraged by the Japanese Fifth Generation project next door) forecast of ever increasing cooperation between AI and database research has remained as elusive as ever. None of us foresaw interoperability (1989*), legacy systems (1989*), middleware (1993*), data mining (1992*), data warehousing (1992*) or Internet (1997*) as future hot topics – although the ability to retrieve information of all kinds from diverse



**Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000**

| Topic Name | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 00 | Total | F3 | L3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AI & Deductive DB | | 3 | | 3 | 1 | 6 | 9 | 5 | 3 | 1 | 9 | 3 | 3 | 3 | 4 | | 1 | | | | | 54 | 1 | |
| DB Machines | | 4 | | | | | 3 | | | 1 | 3 | | | | | | | | | | | 11 | 1.33 | |
| Concurrency Control | 4 | 3 | 3 | 7 | | 3 | | 1 | | | | 3 | | | | 3 | | | | | | 27 | 3.33 | |
| DDBs | 3 | | 3 | 7 | 3 | | 4 | 3 | 1 | | 3 | | 6 | 3 | 3 | | 6 | 9 | 3 | 6 | 4 | 67 | 2 | 4.33 |
| Data Modelling | 9 | 6 | 4 | 11 | 1 | | 6 | 3 | 4 | 3 | 3 | 9 | 6 | | 4 | 3 | | 3 | | 3 | 2 | 80 | 6.33 | 2.67 |
| DB Theories | 3 | 6 | 4 | | 8 | 4 | | 6 | 3 | | | 3 | | 2 | | | | | | | | 39 | 4.33 | |
| DB Trans | | 3 | | | 4 | | | 5 | 6 | 5 | 15 | 3 | 4 | 9 | 3 | 6 | 2 | 3 | | | | 68 | 1 | |
| Query Proc | 3 | 5 | | | 6 | 5 | 12 | 9 | 8 | 8 | 3 | 6 | 3 | 6 | 3 | 6 | 8 | 9 | 12 | 13 | 9 | 134 | 2.67 | 11.33 |
| Views & Der.Rel | | 3 | | | 3 | | | | | 2 | | 1 | | | | | | | | | 1 | 10 | 1 | 0.33 |
| Languages | | | | | | | | | 1 | 3 | 3 | 1 | | 3 | 3 | | | | | | | 14 | | |
| DB Sys Perf | | 2 | | | 3 | 6 | | 3 | 5 | | | 6 | 3 | | 3 | 3 | 3 | | 3 | 6 | 1 | 47 | 0.67 | 3.33 |
| DB Search | | | | | 3 | | | | 1 | 2 | | | | | | 3 | | | 6 | | | 15 | | 2 |
| DB Design | 10 | 5 | 8 | 9 | 7 | 6 | 8 | 3 | 2 | | 3 | | | | 3 | 3 | 1 | 3 | | | 4 | 75 | 7.67 | 1.33 |
| DB Storage | 3 | 6 | 3 | | 4 | 6 | 3 | 3 | 5 | 2 | 6 | 3 | 5 | 8 | 10 | 5 | | 6 | 4 | 8 | 7 | 97 | 4 | 6.33 |
| Multimedia | | 3 | | 1 | 3 | 3 | | 3 | | | | 1 | | | | 3 | | 3 | 2 | | | 22 | 1 | 0.67 |
| Info Retrieval | | | 1 | | | | | | | | | | | 1 | | 3 | 3 | | 3 | 6 | 5 | 22 | 0.33 | 4.67 |
| Object Tech | | | | | | | 3 | | | 7 | 6 | 1 | 4 | 5 | 7 | 6 | 4 | 2 | | | 2 | 47 | | 0.67 |
| Design/Image | | | 1 | 1 | 4 | 3 | | 6 | | 1 | | | | | 3 | | 1 | | | | 1 | 21 | 0.33 | 0.33 |
| Spatial DBs | | | | | | | | | | 1 | 3 | 1 | | | 3 | 2 | 3 | 4 | 4 | | | 21 | | 1.33 |
| Scientific & Stat DBs | | | 4 | 1 | 1 | | | | | 3 | | | | 3 | | | | 1 | | | | 13 | 1.33 | |
| Security | | 2 | | | | | 1 | | | | | | | | | | 1 | | | | | 4 | 0.66 | |
| Time Dimension | | | | | 1 | | 3 | 3 | | 2 | 3 | 3 | 3 | 1 | 1 | | 3 | | 1 | 2 | 3 | 29 | | 2 |
| User Interfaces | 3 | | 3 | 4 | | 3 | | | 3 | 3 | | | | | | | | | | | | 19 | 2 | |
| Active DB | | | | | | | | | | 2 | | 6 | 3 | 3 | 1 | 1 | 1 | | | | | 17 | | |
| Parallel DB | | | | | | | | | | | | | 3 | 3 | 1 | 3 | 3 | 3 | | 3 | 2 | 21 | | 1.67 |
| Data Mining | | | | | | | | | | | | | 3 | | 1 | 6 | 4 | 3 | 10 | 6 | 3 | 36 | | 6.33 |
| C\S & Middleware | | | | | | | | | | | | | | 1 | 1 | | | | | | 1 | 3 | | 0.33 |
| Data Warehousing | | | | | | | | | | | | | 3 | | | | 3 | 3 | 4 | 1 | 2 | 16 | | 2.33 |
| Workflows | | | | | | | | | | | | | | | | | 1 | | | | | 1 | | |
| Internet | | | | | | | | | | | | | | | | | | 3 | | 3 | 6 | 12 | | 3 |
| Agent Tech | | | | | | | | | | | | | | | | | | | | | | 0 | | |
| Total | 38 | 51 | 34 | 44 | 52 | 45 | 52 | 53 | 42 | 46 | 60 | 50 | 49 | 51 | 54 | 56 | 48 | 55 | 52 | 57 | 53 | 1042 | 41 | 54 |

sources all over the world was anticipated [*apparent first reference in a VLDB conference]. As an *aside*, Cobol was predicted to survive the Century as it indeed has done.

In the remainder of this position paper, we shall present the individual contribution of each panelist on the 2020 vision, preceded by a brief analysis database in `the past as represented by the VLDB papers.

## 2. Analysis of the Past: Moderator

At this point it might be interesting to look at the rise and fall of topics in the VLDB conference over the last 20 years (21 including this year). In the table above, I have included only the research papers and grouped them into rough subject categories based on paper titles, choosing from various possible alternatives. The session titles were not always helpful, as they were meant to define sessions, but not necessarily subjects. There was also a question of how many categories I should make. I started with 10 and ended up with about 30, quite arbitrarily as a sort of a canonical basis; but a broader picture can be gleaned from different combinations of related subjects.

The reader might find it interesting to examine the last two columns, F3 the average for the first three years (1980-82) and L3 the average for the last three years (1998-2000), which capture the change over the last 21 years. I make the following observations:

- Some topics are less popular today [AI, Database Machines,…]
- Some important topics had too few papers, but will perhaps come back [Security, Workflows, User Interfaces].
- Some topics are always popular – bread and butter topics [Queries, Transactions, Data modelling, Theories, Performance, ...]
- Some topics are very new [Data mining, Data warehousing, Internet/Web, …] (today's hot topics?)

It seems there were too few papers on User Interfaces, given its importance. Equally surprising is the presence of only one paper on workflows, not to speak of a complete absence

of any paper in the category of agent technology. In fact I have not seen any paper title with the word agent in it.

### 3. S. M. Deen: Agents are Green

Green is a sign for Go. It also implies environmentally friendly, and hence preferable. I argue here provocatively to consider agents as the most appropriate vehicles (and hence preferable) to deliver the goodies in a user-friendly environment from the Internet/Web based complex distributed systems of the future, expected to be commonplace by the year 2020. In addition, I also make some other general points on the future research direction.

The Asilomar Report identifies the forces that will shape database research as being: (i) Internet/Web, (ii) ever complex applications and (iii) advances in hardware. It predicted that multitudes of databases and "trillions of Gizmos" over the Internet/Web would provide special challenges to be overcome with new distributed architectures with ability to provide an interoperable environment. Undoubtedly this is an area of exciting research, which is likely to flourish and bear fruits over the next 20 years on many topics. The anticipated highly distributed environment will in particular need to support open interfaces and complex queries, both of which I think can benefit from the application of agent technology as a means of delivery as claimed below. By complex queries I mean queries from multiple sources, but without a single answer – the "best" answer to be determined by preferences with negotiation and compromises.

Since agents support dynamically changing user-contexts and preferences (with the ability to cooperate negotiate, coordinate and adapt), they should be able integrate multiple (including new) applications that can interface and collaborate with synergy. Such agents should be able to negotiate dynamically on software interfaces forming super applications during the runtime, thus offering scalability and cost effectiveness in an environment where a multitude of databases and "gizmos" are involved. On the other hand the autonomy and use of partial knowledge, which makes agents effective, could also cause severe side-effects, such as poor-decision-making, excessive use of resources, non-convergence and eventually loss of control – all of which could open up new topics for investigation.

Returning to the more general theme, the database research in the first 20 years seems to have culminated in the establishment of the relational model as the central plank, and the single major achievement of the last 20 years is arguably the development of a transaction model, with all its variations.

In the next 20 years, we shall probably be looking at the extensive growth in the middleware technology for handling multitude of information bases distributed over the Internet/Web. Search, navigation, optimization and security will become important topics, often implemented via agent technology, which should also provide platform (middle-ware) independent, dynamically adjustable interfaces and semantic interoperability. How to control emerging group behaviour of agents from their individualised behaviours will be another important topic of study. There will also be richer queries permitting complex retrievals (from disparate sources over the Internet/Web) based on not only constraints but also negotiations and preferences.

Our current concern on distributed data consistency will probably be subsumed within a greater concern on the correctness of behaviour, the consistency of results and timely termination - the problems of complex distributed systems. I suspect User Interfaces will always remain an important topic distinguished by too few VLDB papers.

Finally a highly speculative prediction on new technology: By the year 2020, we should see a paper on quantum processing and/or nano-optimisation in a VLDB conference.

### 4 . A. Jhingran: Porkbellies of the future

It is impossible to predict 20/20, since the changes in the world happen through small transformations connecting the dots towards a larger trend, and interpolation from the first few dots is all but incorrect. However, a 20/10 vision (even better than 20/20!) is much more realistic, and here is my attempt to extrapolate from a few dominant trends.

1. While the PC Installed base will remain stuck in the 600 - 800 million devices, there will be more than 1.2 billion mobile phones by 2010. Data Access will dominate Data Processing.
2. Boundaries of enterprises will transform significantly, with considerable outsourcing of non-strategic functions, including, but not restricted to I/T. Infinite bandwidth will speed up this trend. Consequently, data warehousing will be passe.
3. The network will become much more intelligent, again shifting data and processing away from the enterprise model that we are used to today, but in spite of that Distributed databases and models will not become important. However, "service level agreements" (like QoS) and economic models would become far more important.
4. Value will continue to shift upwards in the food chain, commoditizing database engines (making them like porkbellies -- a thriving business, but most would not care). Application Enabling will be the buzzword for our domain.
5. E-Business will continue to reward those who utilize data for business advantage
6. Moore's Law will continue to work, hence performance will not be a major story

## 5. E. J. Neuhold: Knowledge and Databases – A Vision of the VLDB 2020

Recently I had an interesting discussion with some people from DARPA. One of the challenges to information and knowledge handling and therefore to Very Large Data Bases they see is in a device to directly and continuously dump the contents of our brain, i.e. our personal memory, into some database and retrieve parts again directly into our brain whenever they are needed. Such a device - well known in science fiction literature - would give us perfect memory and as a consequence a much wider base for knowledge, decision making, creativity and wisdom.

Concentrating on the data base aspects of such a device it is very easy to see that the ongoing discussion of the relational model, object oriented approaches, XML databases, or image, video and audio storage mechanisms, even when extended with meta data and semantics will not solve that problem. They are all too far removed from the way our memory is organized, handled, increased, and used. In addition the (partial) reloading of such dumped memory and integrating it with the newly acquired knowledge will compose formidable challenges.

By the way, whenever we dump our knowledge into a computer or other device it will become data/ information and not remain knowledge. We have a tendency to forget that only after we will be able to construct self-aware machines will they be able to say - I know. Without such awareness they will just store data/information and will give back to us that same information or the result of processing these data/information components

## 6. S. Navathe: Databases and DBMSs for Ubiquitous computing in 2020.

Database technology has been evolving for the last 35+ years ever since it was first made available for commercial consumption through DBMS products like IBM's IMS and Honeywell's IDS. In some ways, it has gone a circular route by revisiting concepts that were used in the past by giving them new twists and solving new demands for data. A couple of examples of this circular process stand out that include going from CODASYL network model through relational, back to Object Models which are again graph oriented. The recently popularized XML has the idea of hierarchical organization and access, which was the hallmark of DBMS like IMS and languages like DL/1 based on tree-structured organization of information.

Databases also went through an evolution from centralization to decentralization and distribution to compromise client-server architecture and are coming back to a central theme with data warehousing. From these experiences, it is very obvious that database technology has not necessarily always invented new concepts to meet new challenges, but is likely to revisit, modify and integrate concepts from the past in the next two decades that will address new issues and new applications. In this context, the following four points are particularly relevant:

**A**. Future databases will have to integrate sensors in them attached to external stimuli and they must incorporate the notion of signal processing together with data processing. This will bring with it concomitant problems of sampling, fuzzy matching and pattern understanding that are new as far as basic DB functionality goes. We foresee a very dynamic new research domain that will include EE topics of DSP (digital signal processing) and sensor fusion and domain transforms that aim at handling sampled values from streams for detecting and extracting information, to be coupled with the "standard" notions of query processing and information retrieval.

**B**. Adaptability to users is an area of future development, which will call for models of user behavior based on their psychological underpinnings. User interaction data will have to be archived and processed with technologies such as clustering, classification and data mining at large, which will allow us to understand the users better and adapt the content and display of information in real time.

**C**. The footprint of databases will have to be reduced so that entire databases reside within portable devices, appliances, and gadgets of all types. Rulebases will store rules on user preferences, the understanding of the environment and context under which data need to be processed. Instead of the simple ECA (event-condition-action) rules, we will have a rich repertoire of rules that include economical, psychological and system configuration oriented aspects that must be taken into account before a query is processed or an update is installed in a database.

**D**. It is conceivable that as a "standard" repertoire of functions, database systems may have to offer information requests that require processing of "looks like", "feels like", or "smells like" type of processing. This requires a lot more progress in areas like natural language and speech understanding, understanding of a variety of image formats from a range of application domains. The field of "intelligent databases" seems wide open and is only limited by the types and forms of intelligence we may want to bring into the database functionality.

The exact nature of database management in 2020's is hard to speculate, but it will be a result of some of the pragmatic integration of research from a variety of disciplines together with orders of magnitude advances in storage capacities, bandwidths, and processor speeds with appropriate architectures. Global connectivity and availability of a variety of input/output media will make every individual a database client and every electrical/electronic device a database processor by 2020!

### 7. G. Wiederhold: Will database research really support decision-making?

Although database systems have been promoted by their developers as supporting decision- makers we find very few instances where databases are used directly by the decision-makers. The primary computer-based tool used today by actual decision-makers is the spreadsheet, which appears to provide satisfactory interaction. Even when spreadsheet data and formulas have been initialized by an intermediary specialist, the decision-maker can easily plug in parameters to evaluate alternate futures, an essential aspect of planning.

In planning, the decision-maker does not only need data about the past, as provided by databases and data warehouses, but also information that provides projections into possible futures. The future is in part determined by actions the decision-maker can take, and in part by reactions and independent events that may occur in the world. Is the database community interested in providing services to meet such requirements? I would like to assume so, but it will require an expansion of the database researchers' mindsets.

Many concepts from databases will remain valuable, such as schema-driven execution, query languages, distributed access, temporal functions, attached procedures, and caching to gain performance.. But additional concepts will be needed and must be formalized and integrated to build a new generation of information system. Predictions are often based on substantial computations, which may be best supplied by remote processors. Many intermediate results will have associated uncertainties, and these must be aggregated with their data. There will not be a single correct timeline, representing the past, but a bush of future alternatives. Queries must be able to return multiple sets of values for a future point-in-time, and these value sets must be labeled with the actions and assumptions that led to them.

The space needed for all results obtainable for all future situations is immense, and cannot be stored, so that it becomes essential to provide execution-time linkages to computations. The results of these computations must be seamlessly integrated with results derived from databases. Results from planning sciences can help in combining uncertainties and pruning unlikely or low-value alternatives to keep the information volume presented to a decision-maker modest.

Set-based algebras from databases may be combined with ranking schemes from web-based searches. Interpolation is needed to provide values at arbitrary future points in time. The simplest approach is to assume a linear function to estimate the intermediate values, although other more involved approaches can sometimes be required.

The benefit of being able to interpolate for missing values is not limited to the predictive system capabilities. Anywhere where functions are naturally continuous interpolation has benefits, since data in a database are always associated with discrete points, or are means over an interval bounded by discrete points. For instance, observations at temporal and spatial points rarely match the query. We may want to the temperature and wind speed and direction at Gizeh, but have data only for major cities, say Cairo, Port Said, Alexandria, etc. Interpolation is also important in computer-aided design, to find optimal materials satisfying stress, flex, lifetime, and environmental conditions. Satellite data are not directly linked to points known to geographers or politicians, and require interpolation.

Architecturally there are a number of questions to be resolved. How much of these support functionalities belong inside the database systems and which are best served by external services? Where should caching take place? How will schemas handle computed elements and their identification? The research issues are broad, and it will be interesting to see where new research efforts will take place.

## Summary and Remarks

In addition to the Introduction and Review, this position paper includes contributions from all Panel members expressing their differing individual visions of the database research of the future, using the year 2020 as target.

In this Deen has focused on requirements of complex distributed systems with agent technology playing a significant role, while Jhingran depicted the operational environment in terms of both the infrastructure facility and core technology. Neuhold considered future databases as holders of not only information but also real knowledge (including human knowledge) as data with connection to the human brain. Navathe has explored the role of databases in ubiquitous computing which in the future should offer a wide range of facilities including individualized *touchy*, *feely* and even *smelly* services. On the other hand, Wiederhold pondered on research needed, and likely to be carried out in the next two decades, to extend the database facility to the actual decision--makers.

These presentations were not meant to offer a comprehensive coverage of future directions of database research, nor to provide mutually consistent and non-controversial visions, but to provoke and stimulate discussion at the Panel meeting.

**Acknowledgement**:

Thanks to Thomas Neligwa of Keele for struggling painstakingly with MS Word and the VLDB page format, and finally getting everything neatly into five pages.