A Database Striptease or How to Manage Your Personal Databases

Martin Kersten CWI, Netherlands, Gerhard Weikum Univ. Saarland, Germany
Michael Franklin U.C. Berkeley , USA Daniel Keim Univ. Konstanz, Germany
Alex Buchmann T.U. Darmstadt, Germany Surajit Chaudhuri Microsoft Research, USA

1 Setting the stage

As a thief in the night, the database management problem has been entering everyone's life. To realize its presence it suffices to sit down and tally the electronic datasources crucial for survival in our modern society.

Mister Average has several dozen of databases to take care of. On his body he carries at least half a dozen cards with tiny databases of personal data which have been constructed to federate with large datasources through special devices. Many of these database have a high economic value or are plain life-savers. The PDA in his hand provides a cache for several datasouces on his laptop or PC, including the address database, personal financing, and project dossiers. A sizeable MP3 database may be included, or stored in a separate compartment carried in his pockets. The wristwatch seems like a single function device, but the latest incarnations have the processing and storage capacity to take over part of the PDAs functionality. A telephone complements his luggage with at least a database of telephone numbers, accounting database for its use, message database and even a multimedia database of pictures captured with a 'dime-sized' camera.

In his car, mr Average should manage his audio database. They may be as small as the 6-record database with preset channels to listen too or as complicated as a large MP3 cache with preselected play lists (and DVDs for his kids in the rear). His car-navigation system currently only stores the latest trips, but could be extended to establish a liaison with his PC to automate the administration for cost-claims. Here too, we might find an integrated telephone with (a separately managed) phone book. Waiting in a traffic jam the car soon becomes a manage-by-the-minute office space, where email is handled before reaching the office.

His home setting becomes the playground of Ambient devices, which are sensitive, adaptive, and responsive to the presence of people. They are designed to improve the quality of life by creating a desired atmosphere and functionality via intelligent, personalised inter-connected systems which blend with the physical background. The devices keep extensive historical records of highly sensitive personal (physical) date. Yet, the people are not aware of their existence nor the extend of this big brother administrative behaviour.

2 The plot

As long as datasources are independent, devices are never replaced, nor new devices enter our realm of existence, we will survive easily in the digital jungle. However, life runs a different course. Each time we meet a new person, we may have to synchronize several databases with his address information. The limitations of the human brain to cope with the information overload calls upon better support to 'remember' where, what and when has been accumulated in the fabric of datasources making up our environment. Buying a new PDA (from a different brand) surely means a re-organization and possibly retyping the content of the database. Even when devices are linked into a communication network data interoperability may not be achieved, e.g. the GPS system of your car navigation system can not be used by your digital camera to record the location. It is not even possible to perform a temporal join over the GPS record and your digital camera record a posteriori.

So how does Mr. Average search his wealth of personal data, spread across a plethora of devices, for some specific piece of information that he desperately needs? Like most humans, Mr. Average does not remember detailed facts about events long ago but merely recalls vague associations. For example, he may be looking for a photo of the nice Dutch family that

Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003

he met when climbing in the Bavarian Alps in summer 2001. Or was it actually on a glacier hike in Switzerland that he did in 2000? And didn't he exchange email with them later, or even looked at the family's photo album on their Web homepage? Queries of this kind operate on a variety of data that range from structured (e.g., email address books or personal expenses and finances) to highly unstructured and poorly organized data (e.g., photo album annotations and email folders), and, of course, also multimedia content. Answering them requires coping with vagueness and ambiguities in text, data, and the query itself, and it calls for approximate joins across temporal as well as spatial dimensions. The search result should be customized to the specific habits of Mr. Average, in terms of his personal style of phrasing queries, his typical way of annotating photos, and his general background and life experience (to the extent that is known to at least one of his many database devices). This calls for ontological background knowledge that can be exploited for more effective query processing. However, this kind of reasoning must take time and space perspectives into account; for example, when Mr. Average asks about a picture of the web that he was so fascinated about when he was a child (30 years ago), the system should know that he means a real spider web. Finally, another aspect to consider is that Mr. Average, like most humans, is very lazy and sloppy in organizing his personal information. So support is crucial for automatic annotation, classification, and organization of email, photos, etc., which in turn would help the system for executing queries efficiently and effectively.

3 The play

Taken in isolation, each application appears as a trivial task and developers will not be inclined to consider a DBMS the right approach to manage a list of a few tens of records. The investment is too high and the macro benefits (interoperability and evolution) are unclear. At the same time, to secure product lines and enable interoperability between products of a single supplier, the application developers are forced to rethink transaction management, resource optimization, and query processing.

It is up to the database community to avoid this re-invention of the wheel by timely provision of the necessary technology, algorithms, and software components. Unfortunately, the VLDB research community too considers each database management task as too simplistic, thereby missing the big picture.

The panelists are challenged to comment on the opportunities, challenges, pitfalls, and laboratory progress on database technology for the personal databases.

- *Organic Databases* Future database software solutions are not measured by their compliance of the SQL or XQUERY standard, nor their performance on TB stores, but along a completely different dimension. Namely, can we develop an "Organic" DBMS which can be embedded in a wide collection of hardware appliances and provide an autonomous, self-descriptive, self-organizing, self-repairing, self-aware, secure and stable data management functionality to its environment

- *Searching Personal Information* The way we support querying should be drastically improved. The assumption that any query is ran against a database system without a priori knowledge, nor contextual information requires rethinking the algorithms, query answer caching and cost-models.

- *DBMS Product Evolution* Old software technology is here to stay, it can and will adapt. Given the trend in computing resources, the small devices become also the playground of the products already on the market. This means an evolutionary road seems sufficient.

- *P2P and Stream-based Architectures* The vast array of devices embedded in our day-to-day environments will be best addressed, accessed, and manipulated using declarative query processing-like techniques with roots in database technology. The challenges are numerous, and include: resource discovery, result fusion, fault tolerance, non-traditional optimization, spatio-temporal processing, caching, archiving and purging, and privacy.

- *Gossips-based communication* Interoperability requires architectures based on event messaging infrastructures. In particular, applications set a channel to gossip amongst one-others in a mutual agreed language. We envision two kinds: the local gossip and other guys have combined P2P with pub/sub for large scale.

- *Visual Feedback* Standard text-oriented query interfaces are certainly not useful for querying and exploring personal databases. Pictures worth a 100 Megabyte of records are needed to explore the databases and together with direct visual interaction they will help the user to find the needle in the haystack.