

Big Science Data Coming to SQL Databases

Multi-dimensional sensor, image and simulation output and statistics data make up most of the 'big data' in science and engineering. In June 2014, it has been decided to extend the SQL database language with massive multi-dimensional arrays. Initiated by Prof Peter Baumann from Jacobs University, Bremen, Germany, work has commenced on the forthcoming standard named SQL/MDA.

SQL has been tremendously successful in running any-size databases in business and administration. However, 'big data' in science is structured differently. Instead of simple tables, it often consists of multi-dimensional 'data cubes'. In geo sciences, for example, this encompasses 1-dimensional sensor data, 2-D satellite imagery, 3-D x/y/t image time series as well as x/y/z geophysical voxel data, and 4-D x/y/z/t weather data. In life sciences, there is laserscan microscopy and brain scans. And this can grow as large as simulations of the whole universe when it comes to astrophysics.

Rasdaman

But SQL is not able to find, filter, and process such arrays, and consequently arrays are currently largely maintained outside databases. Recognising this shortcoming, Peter Baumann, professor of computer science at Jacobs University Bremen, and his group have long been researching ways to extend SQL appropriately. The rasdaman system which the group has established effectively has coined a new technology, Array Databases. In a recent technology demonstration, more than 1,000 computers have collaborated in a cloud to jointly compute the result of one single database query. This 'distributed query processing' means a massive speedup, and research challenges on multi-Petabyte data cubes can be answered that hitherto were unsolvable.

Meantime, international data centres use this tool to allow scientists to gain unanticipated insights into their spatio-temporal data cubes, and rasdaman installations can be found at NASA, ESA, British Geological Survey, Plymouth Marine Laboratory, Deutscher Wetterdienst, and many more.

At a recent meeting in Beijing, all national bodies participating in the SQL working group of ISO unanimously agreed on the importance of arrays in SQL. Following thorough assessment of all options available, the group accepted Baumann's proposal for further elaboration. The new standard will be named ISO 9075 SQL/MDA which stands for 'Multi-Dimensional Arrays'.