From Terabytes to Petabytes: Computer Scientists Develop New Hybrid Database System

August 26 2009

(PhysOrg.com) -- As the amounts of data being stored by databases around the world enters the realm of the petabyte (the amount of data stored in a mile-high stack of CD-ROM disks), efficient data management is becoming more and more important. Now computer scientists at Yale University have developed a new database system by combining the best features of multiple approaches to create an open source hybrid system called HadoopDB.

Traditional approaches to managing data at this scale typically fall into one of two categories. The first includes parallel database management systems (DBMS), which are good at working with structured data that contain, for instance, tables with trillions of rows of data. The second includes the kind of approach taken by MapReduce, the software framework used by Google to search data contained on the Web, which gives the user more control over how the data is retrieved.

“In essence, HadoopDB is a hybrid of MapReduce and parallel DBMS technologies,” said Daniel Abadi, assistant professor of computer science at Yale and one of the system designers. “It’s designed to take the best features of both worlds. We get the performance of parallel database systems with the scalability and ease of use of MapReduce.”

HadoopDB was announced on Abadi’s blog last month. Yale graduate students and co-creators Azza Abouzeid and Kamil Bajda-Pawlikowski

will present more in-depth details of the new system at the VLDB conference in Lyon, France on August 27. They will also present results of a detailed performance analysis they conducted with Abadi, Avi Silberschatz, chair of computer science at Yale, and Alexander Rasin of Brown University. The team will demonstrate the system performance on a range of representative queries at the conference, both on structured and unstructured data, and will outline HadoopDB's characteristics along the run-time performance, loading time, fault tolerance and scalability dimensions.

With the huge amounts of data being collected and used in today's databases - from consumer information used by retail chains to improve buying experiences and reduce customer churn to financial information being collected by banks to reduce risk and avoid another catastrophic financial collapse- being able to store and analyze such vast amounts of data will only continue to grow in importance, Abadi said.

HadoopDB reduces the time it takes to perform some typical tasks from days to hours, making more complicated analysis possible - the kind that could be used to find patterns in the stock market, earthquakes, consumer behavior and even outbreaks, Abadi said. "People have all this data, but they're not using it in the most efficient or useful way."

Provided by Yale University (news : web)