Relational Join Using Map-Reduce

Professor Jeffrey Ullman
Computer Science Department
Stanford University

Date: 19 May 2011 (Thursday)
Time: 4.30pm – 5.30pm
Venue: SPMS-LT5, SPMS-03-08
School of Physical and Mathematical Sciences

We introduce the environment of map-reduce systems like Hadoop, including distributed file systems and SQL-like systems built on top of Hadoop. There is a simple algorithm for implementing the usual 2-way join in map-reduce, which we shall review. Then we consider the problem of joining more than two relations in one map-reduce round. We see that there is a nonlinear optimization problem that must be solved to do the task most efficiently. Sometimes the multiway join is in fact more efficient than a cascade of 2-way joins, although in many cases it is not. We then concentrate on one important case where the multiway join, properly implemented, is always a win: the case of star joins (analytic queries).

Speaker Biography

Professor Jeffrey Ullman is the Stanford W. Ascherman Professor of Computer Science (Emeritus), at Stanford University. He has published 19 books and 187 research papers, and made significant contributions to theoretical computer science and database, as well as many other areas. He is a member of National Academy of Engineering and an ACM fellow. He has won a number of awards, including, Knuth Prize, SIGMOD Contributions Award and Test-of-Time Award, IEEE von Neumann Medal, ACM Karl V. Karlstrom Outstanding Educator Award.

Host: Division of Mathematical Sciences, School of Physical and Mathematical Sciences