MapReduce-based Spatial Join

Keywords: Scientific Data Management, Spatial Indexes, Neuroscience

Context: Scientists in all kinds of disciplines like biology, chemistry, physics etc. produce vast amounts of data through experimentation and simulation. The amounts of data produced are already so big that they can barely be managed. And the problem is certain to get worse as the volume of scientific data doubles every year. In the DIAS laboratory we are working on next generation data management tools and techniques able to manage tomorrow’s scientific data.

We work with neuroscientists in the Blue Brain Project (http://bluebrain.epfl.ch) to manage the vast amounts of data they produce. Their research, modeling and simulating a fraction of the rat brain, already produces gigabytes of data. With the recent upgrade of their computing infrastructure (IBM Blue Gene/P), the volume of data will soon be in the order of terabytes.

Current solutions are inadequate to manage this data volume and we are thus investigating new methods to index and store it in order to provide efficient access. A particular problem we are currently investigating is where different neurons touch another. While it is simple to compute this for several thousand neurons, we will have to do this for brain models featuring several millions or even billions of neurons.

Project: Before simulation, the neuroscientists build a small model of the brain containing the two basic primitives, neurons and synapses (connections between neurons). In the first step several thousand neurons are put together to build a model. Each neuron is modeled with thousands of cylinders and is given a shape (i.e., each cylinder has a particular location) determined during the in-vitro analysis of rat brain tissue. In the second step, synapses are placed at the places where neurons intersect each other and where impulses may leap over during the simulation.

Placing the synapses is a computationally very complex problem, but fortunately is equivalent to a well-known and extensively studied data management problem, the spatial join [1]. Current solutions, however, are not efficient when performing this operation on millions and billions of neurons and the goal of this project therefore is to implement a scalable spatial join on top of MapReduce [2] so that we can solve the problem on ever bigger brain models massively parallel.
Milestones: 1. Get acquainted with the spatial join methods developed and described in related work.
2. Get acquainted with the Hadoop [3] framework as an implementation of MapReduce
3. Implement the spatial join (PBSM [4] approach or similar) on Hadoop. PBSM works in two phases: 1) it assigns each object to a grid cell and b) compares all objects in a cell pairwise. For an implementation of PBSM on Hadoop, the following steps are necessary:
   a. Take as input the two data sets A & B
   b. In the map phase: assign each object of A and B to the respective grid cells
   c. In the reduce phase:
      i. ensure all key-value pairs belonging to the same cell end up on the same reducer
      ii. compare all key-value pairs in the same grid cell with an efficient in-memory spatial join (available)
   d. as part of the reduce phase (or thereafter) deduplicate results
4. Use advanced features of Hadoop to optimize the join, e.g., combiner
5. Extensively and thoroughly evaluate performance and compare to a base-line approach that does not use Hadoop

Knowledge: C/C++

Supervisor: Prof. Anastasia Ailamaki (anastasia.ailamaki@epfl.ch)
Responsible collaborator(s): Thomas Heinis (thomas.heinis@epfl.ch), Farhan Tauheed (farhan.tauheed@epfl.ch)
Duration: 2 months

References