Title: On-the-Fly Processing of Big, Distributed, Streaming Data
Author: Assaf Schuster

Abstract

More and more tasks require efficient processing of continuous queries over scalable, distributed data streams. Examples include optimizing systems using their operational log history, mining sentiments using sets of crawlers, and data fusion over heterogeneous sensor networks.

However, distributed mining and/or monitoring of global behaviors can be prohibitively difficult. The naïve solution which sends all data to a central location mandates extremely high communication volume, thus incurring unbearable overheads in terms of resources and energy. Furthermore, such solutions require expensive powerful central platform, while data transmission may violate privacy rules. An attempt to enhance the naïve solution by periodically polling aggregates is bound to fail, exposing a vicious tradeoff between communication and latency.

Given a continuous global query, the solution proposed in the talk is to generate filters, called safe zones, to be applied locally at each data stream. Essentially, the safe zones represent geometric constraints which, until violated by at least one of the sources, guarantee that a global property holds. In other words, the safe zones allow for constructive quiescence: there is no need for any of the data sources to transmit anything as long as all constraints are held with the local data confined to the local safe zone. The typically-rare violations are handled immediately, thus the latency for discovering global conditions is negligible.

The safe zones approach makes the overall system implementation, as well as its operation, much simpler and cheaper. The saving, in terms of communication volume, can reach many orders of magnitude. The talk will describe a general approach for compiling efficient safe zones for many tasks and system configurations.