Interactive Data Processing at Massive Scale

Magda Balazinska
(U. Washington)

Abstract

The ability to analyze massive-scale datasets has become a critical requirement for industry and sciences alike. Because of the magnitude of the data involved in today's applications, users are increasingly turning toward parallel data processing systems running in shared-nothing clusters. These systems provide efficient query processing facilities, but the magnitude of input data sets still causes most queries to take from tens of minutes to several hours to complete. At this scale, users need more than efficient processing. They also need effective tools for managing their queries at runtime including accurate, time-based progress indicators, the ability to suspend and resume queries, the ability to see representative partial results, intra-query fault-tolerance, and agile query scheduling and resource management mechanisms. All this without too much runtime overhead.

In this talk, we will present our vision and preliminary results for a massive-scale data management system that enables interactive data processing at massive scale by providing the features outlined above.

We will also touch on our long-term vision to build science-oriented database management services in the cloud.

This work is part of the Nuage project at the University of Washington: http://nuage.cs.washington.edu