Grand System Research Challenge Problem:

Monitoring the Data Tsunami

Johannes Gehrke
Department of Computer Science
Cornell University

In 2020, we will live in a world that is networked to unprecedented scale, and computer networks have become more pervasive in scope and mission-critical to businesses, scientific endeavors and other computing applications. But at the same time, we face an increasing problem of network instability and unreliability. Anyone who uses the web is familiar with such problems as servers that respond sluggishly (or not at all), network connections with erratic throughput, regional disruptions occasioned by accidental events, political events, or distributed denial of service attacks. The networked world of the future will have massive scale but, absent some significant technical advances, will also be pervasively unreliable.

At the same time a revolution at the application level is happening, with types of applications that are heavily network-centric, and are best characterized as a *monitoring applications*. Every device and appliance will have computing and communication capabilities, and smart sensor networks will be deployed widely for measurement, detection, and monitoring applications. Senior citizens will wear continuous measurement devices that will remind them of their medication, monitor blood-sugar and cholesterol levels, and their physicians will use this information to adjust medical treatments in direct dialog with the monitoring outfit. Soldiers will be outfitted with arrays of environmental and physical sensing systems, and mission planners will combine this information with a flood of information from other forms of sensors to create situational maps in real-time with astonishing detail. The regional airport will track supplies and replacement parts in real-time, coordinating this information to anticipate the need for parts and order them just as needed. In the restructured electric power grid, power producers and consumers will monitor grid status, demand and production to match production to demand and set pricing. Infrastructure for such monitoring applications will generate floods of data of unseen scale.

Existing architectures for monitoring applications are inadequately scalable, overly constraining, and vastly unreliable. For example, today's sensor networks lack flexibility because sensors are treated as remote peripherals from which data can only be extracted in a predefined way for transmission to central storage. Existing peer-to-peer systems show some promise to being able to deal with decentralized and autonomous participants, but so far they can handle only very simple storage tasks. Centralized architectures do not scale to large monitoring applications; they would overwhelm both the network and the centralized storage system, and would end up with a single point of failure.

We need technology to unite the seemingly conflicting requirements of scalability, reliability, and ease of programming and user interaction in monitoring the physical world. We need a new distributed data management, data mining, and networking infrastructure that scales with the growth of sensor, device, and actuator interconnectivity and computational power over the next decades, technology that enables us to respond to the impending data tsunami.

Biographical Sketch for
Johannes Gehrke
Cornell University

February 27, 2002

Johannes Gehrke is an Assistant Professor in the Department of Computer Science at Cornell University. Johannes received his PhD from the University of Wisconsin-Madison in 1999, where his graduate research was supported by an IBM Fellowship. Johannes joined Cornell University in August 1999 as an Assistant Professor. Johannes received an IBM Faculty Development Award both in 2000 and 2001, the Cornell College of Engineering James and Mary Tien Excellence in Teaching Award in 2001, and a National Science Foundation CAREER Award in 2002.

At Cornell, Johannes leads the Himalaya Data Mining Project, the Amazon Stream Processing Project, and the Cougar Sensor Database Project. The Himalaya Project develops data mining technology for new applications, as well as techniques to make the resulting data mining models better understandable for the user. The Amazon Project develops query processing techniques and a system architecture for long-running queries over high-speed data streams. The Cougar Project builds a distributed data management infrastructure for sensor networks that scales with the growth of computational power and interconnectivity of future sensor networks.

Johannes has published numerous papers on data mining and database systems, and he has given several tutorials on data mining at conferences and on Wall Street. Johannes is the co-author of the textbook "Database Management Systems (Second Edition)", published by McGrawHill in 1999.

Selected publications:

Johannes E. Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, and Wei-Yin Loh. BOAT – Optimistic Decision Tree Construction. In *Proceedings of the 1999 ACM Sigmod International Conference on Management of Data*, Philadelphia, Pennsylvania, 1999.

Johannes E. Gehrke, Raghu Ramakrishnan and Venkatesh Ganti. RainForest – A Framework for Fast Decision Tree Construction of Large Datasets. *Data Mining and Knowledge Discovery* 4(2/3): 127-162 (2000)

Venkatesh Ganti, J. E. Gehrke, and Raghu Ramakrishnan. DEMON: Mining and Monitoring Evolving Data. *IEEE Transactions on Knowledge and Data Engineering* Vol. 13, No. 1, January/February 2001, pages 50-63.

Johannes E. Gehrke, Flip Korn, and Divesh Srivastava. On Computing Correlated Aggregates Over Continual Data Streams. In *Proceedings of the 2001 ACM Sigmod International Conference on Management of Data*, Santa Barbara, California, May 2001.

Alin Dobra, Minos Garofalakis, J. E. Gehrke, and Rajeev Rastogi. "Processing Complex Aggregate Queries over Data Streams". To appear in *Proceedings of the 2002 ACM Sigmod International Conference on Management of Data (SIGMOD 2002)*, Madison, Wisconsin, June 2002.