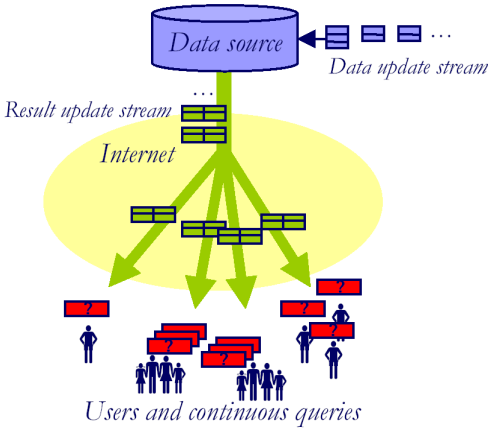
Recent Developments in **CAREER: Techniques and Applications of Derived Data Maintenance**

IIS-0238386 PI: Jun Yang (Duke University) January 29, 2007

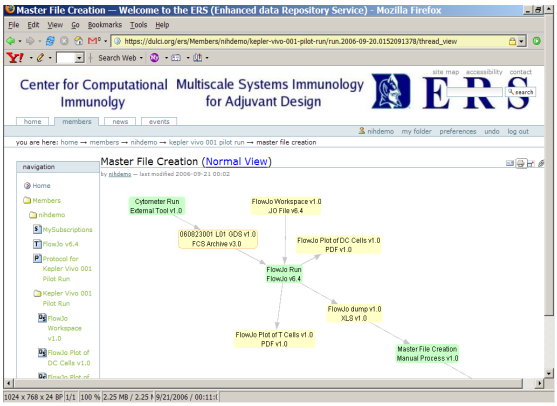
Background. The focus of this CAREER project is on *derived data maintenance*. Derived data is the result of transforming *base data*. The use of derived data to facilitate access to base data is a recurring technique in many areas of computer science. Maintenance of derived data, whenever base data is updated, is a fundamental problem with many applications. It is also an evolving problem—existing techniques are constantly challenged by the explosive growth in data volume and number of data producers and consumers, and by increasing diversity in data formats.

Application in Continuously Query Systems. Traditional database systems allow users to ask questions over the current snapshot of the data. Today, more and more applications require the ability to monitor data that is constantly evolving—for example, financial market information, traffic and weather conditions, patient vital signs, battlefield situations, etc. Periodic polling of data sources is inefficient and may miss important updates. A powerful model base on *continuous queries* has emerged, where queries continue to generate new results (or changes to results) and deliver them to the users, whenever new data updates stream in. Continuous query processing can be regarded as derived data maintenance. We have made significant contributions in improving the efficiency of continuous query systems, making them more scalable when the rate of change is high [11], and when there are a large number of users and queries [3, 10]. We are also investigating continuous queries over a wide-area network, with applications in monitoring large networks [8] and Internet-based subscription services [4] such as Google Alert.



Application in Environmental Monitoring. With a team of ecologists, statisticians and engineers at Duke and Northern Arizona University, we are building a wireless sensor network in the Duke Forest [7]. The sensor network will help researchers better understand how the growth, survival and reproduction of forest trees are influenced by changes in climate, atmospheric carbon dioxide, disturbances, and other environmental factors that can fluctuate rapidly. The task of the sensor network is to continuously provide an update-to-date view of the environment derived from sensor readings. We have developed a collection of novel derived data maintenance techniques [1, 2, 5, 6, 9] in this setting, aimed at reducing energy consumption—often the most critical factor in prolonging the lifetime of the battery-powered sensor nodes.

Application in Immunology Research. We are collaborating with computational immunologists at Duke on a system for tracking lineage, dependency, and versioning of derived datasets in computational biology workflows. Today, biomedical research is increasingly done *in silico*. Many computational tasks (such as data extraction, analysis, and visualization) involve complex workflows that begin with multiple sources of data, transform them in multiple steps using software tools, and generate results. In a collaborative environment, results produced by one user's workflow may be the input to another user's workflow. When any input changes, or any software tool used in the workflow has a bug fix or new version, results may need to be regenerated. We are building a database for tracking dependency and version information, as well as a user-friendly Web interface for entering and accessing such data. Users automatically receive notifications when relevant inputs or tools have been updated. A number of our collaborators have begun using this system, and we hope to make it available to general public after more testing.



Recent Publications.

1. Adam Silberstein, Rebecca Braynard, Gregory Filpus, Gavino Puggioni, Alan Gelfand, Kamesh Munagala, and Jun Yang. "Data-Driven Processing in Sensor Networks." In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (CIDR '07)*, Asilomar, California, USA, January 2007.
2. Adam Silberstein and Jun Yang. "Multiple Aggregation for In-Network Control of Sensors." *In Proceedings of the 23rd International Conference on Data Engineering (ICDE '07)*, Istanbul, Turkey, April 2007.
3. Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu. "Scalable Continuous Query Processing by Tracking Hotspots." *In Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06)*, Seoul, Korea, September 2006.
4. Badrish Chandramouli, Junyi Xie, and Jun Yang. "On the Database/Network Interface in Large-Scale Publish/Subscribe Systems." In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*, Chicago, Illinois, USA, June 2006.
5. Adam Silberstein, Rebecca Braynard, and Jun Yang. "Constraint-Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks." In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*, Chicago, Illinois, USA, June 2006.
6. Adam Silberstein, Kamesh Munagala, and Jun Yang. "Energy-Efficient Monitoring of Extreme Values in Sensor Networks." In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*, Chicago, Illinois, USA, June 2006.
7. Paul G. Flikkema, Pankaj K. Agarwal, James S. Clark, Carla Schlatter Ellis, Alan Gelfand, Kamesh Munagala, and Jun Yang. "Model-Driven Dynamic Control of Embedded Wireless Sensor Networks." In *Proceedings of the 6th International Conference on Computational Science (ICCS '06)*, Reading, United Kingdom, May 2006.
8. Badrish Chandramouli, Jun Yang, and Amin Vahdat. "Distributed Network Querying with Bounded Approximate Caching." In *Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA '06)*, Singapore, April 2006.
9. Adam Silberstein, Rebecca Braynard, Carla Ellis, Kamesh Munagala, and Jun Yang. "A Sampling-Based Approach to Optimizing Top-k Queries in Sensor Networks." In *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, Atlanta, Georgia, USA, April 2006.
10. Pankaj K. Agarwal, Junyi Xie, Jun Yang, and Hai Yu. "Monitoring Continuous Band-Join Queries over Dynamic Data." In Proceedings of the 16th Annual International Symposium on Algorithms and Computation (ISAAC '05), Sanya, Hainan, China, December 2005.
11. Junyi Xie, Jun Yang, and Yuguo Chen. "On Joining and Caching Stochastic Streams." In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, Baltimore, Maryland, USA, June 2005.

For earlier publications resulted from this project, please visit the project Web page at http://www.cs.duke.edu/dbgroup/ddm/.