Large-scale data analysis is a key driver not just of business decisions and application logic, but also major recent innovations in computer science. However, the data analytics landscape is facing constant disruption. Trends such as geo-distribution of datasets, with the associated regulations on data movement, the emergence of analytics-as-a-service, and the popularity of new compute substrates such as serverless computing, edge computing, and spot markets are stretching the designs of existing data analytics stacks beyond their original targets. This has left even state-of-the-art analytics stacks incapable of supporting modern-day analytics.

I will describe some of the systems my group has developed that highlight general principles and building blocks for fast, flexible analytics in the face of such disruption. Carbyne is a new cluster management system for analytics-as-a-service that leverages cross-layer design and implicit cross-analytics coordination to ensure efficient, fast, and resource-fair analytics. Clarinet is a new multi-query optimizer for geo-distributed analytics that is built atop explicit cross-query coordination and delayed query plan binding. QOOP refactors the interfaces between, and the responsibilities of, query optimizers, execution engines, and schedulers to enable bottom-up feedback and query replanning in the face of volatility in the resources available for analytics which might arise, for example, in spot markets. I will conclude with an overview of our current work on Whiz, a new, completely re-envisioned data analytics stack. Whiz elevates intermediate data as a first-class entity, enabling the highly-adaptive and general “data-driven computation” paradigm for accelerating batch stream and graph analytics.

For more information or questions:
Karen Lindenfelser, 8-6716, karen@ece.cmu.edu
http://www.pdl.cmu.edu/SDI/