Postdoc Position in Big Data Analytics

Database and Analytics Research Group
École Polytechnique, France

Contact:	Prof. Yanlei Diao (École Polytechnique, yanlei.diao@polytechnique.edu) http://www.lix.polytechnique.fr/~yanlei.diao/
Duration:	12-24 months, available immediately
Application:	Submit a CV, three names of references, and 1-2 research papers to Prof. Diao by email

The Database and Analytics Research Group at École Polytechnique is seeking a postdoctoral fellow in the area of Big Data analytics and optimization. The appointment would be for one year, with possible renewal for the second year. The postdoctoral fellow will find an active and collaborative environment at École Polytechnique, with world-renowned researchers in database systems, database theory, statistics, and machine learning.

École Polytechnique is a French public institution of higher education and research, located in Palaiseau 45 minutes southwest of Paris. It is considered the most prestigious engineering school in France, with well-known educational programs in science and engineering. Among its alumni are three Nobel prize winners, one Fields Medalist, three Presidents of France, and many CEOs of French and international companies.

Topic: A Next-Generation Unified Data Analytics Optimizer

Today's big data analytics systems are best effort only: despite the wide adoption, they still lack the ability to take user budgetary constraints and performance goals, and automatically configure an analytic job to achieve those goals. In this project, we aim to take a step further towards building a *next-generation unified data analytics optimizer* that takes as input a user analytic task in the form of a dataflow program (which subsumes traditional SQL queries, machine learning tasks, graph analytics, etc.), as well as a set of user budgetary constraints and performance goals, and produces as output a cloud instance and runtime parameters of the job that best meet the user objectives. The data analytics optimizer has the potential to offer crucial benefits to users and cloud service providers, such as finding an an optimal choice that best meets the user objectives and allowing the cloud service provider to best utilize available resources to support a larger user population.

As a first step towards this vision, this project focuses on the design of an optimizer that will enable automatic job configuration based on user objectives in the setting of dedicated clusters. The project will explore cost modeling and optimization strategies for dataflow programs that are treated as blackboxes (e.g., a machine learning program, or a UDF embedded in a SQL query), and further leverage semantic and logical properties of query plans, if present, to achieve improved performance.

Desired background for the postdoc position: Applicants must have a Ph.D. in Computer Science with a background in database systems and Big Data systems. Substantial knowledge of query processing, query optimization, and distributed Big Data systems such as Spark is expected from qualified candidates. One or two research papers on related topics, as well as experience of implementation and experimentation with Big Data systems, are strongly preferred. In addition, knowledge of statistics and machine learning will be very helpful to the project.