Dissertation Defense

From Droplets to Cloud: Towards Privacy-Preserving Integration of Distributed Heterogeneous Data

Pawel Jurczyk
Emory University

Abstract: With the trend of cloud computing, data and computing are moved away from desktop and are instead provided as a service from the cloud. Data-as-a-service enables access to a wealth of data across distributed and heterogeneous data sources in the cloud. It remains a challenge, however, to ensure the privacy, interoperability, and scalability for such services.

We designed and developed DObjects, a general-purpose P2P-based query and data operations infrastructure that can be deployed in the cloud and provides access to heterogeneous data sources. The system builds on top of a distributed mediator-wrapper architecture where individual system nodes serve as mediators and/or wrappers and interact with each other in a P2P fashion what guarantees good scalability. As an analogy, the system nodes can be considered as droplets, small elements that provide similar functionality in the cloud. Just as thousands or millions of droplets form a single drop in nature, in cloud computing, groups of droplets that provide similar functionality can form a micro-cloud. Micro-clouds are an integral part of the whole cloud computing system and can provide specific services to users.

The dissertation also discusses the novel dynamic query execution engine within the data query infrastructure that dynamically adapts to network and node (or droplet) conditions. The query processing is capable of fully benefiting from all the distributed resources to minimize the query response time and maximize system throughput. In addition to leveraging the traditional distributed query optimization techniques, the (sub)queries are deployed and executed on droplets in a dynamic and iterative manner in order to guarantee the best reaction to network and resource dynamics.

Finally, the dissertation presents an extension to the basic DObjects model that enables access to private data that is distributed and needs anonymization. The extension enables droplets to form virtual groups in order to addresses two important privacy issues for the sensitive data: privacy of data subjects and confidentiality of data providers. The dissertation discusses decentralized protocols that enable data sharing for horizontally partitioned databases given these constraints. These protocols can be run by the groups of droplets. Concretely, given a query spanning multiple databases, the query results do not contain individually identifiable information. In addition, institutions do not reveal their databases to each other apart from the query results.

Tuesday, December 1, 2009, 3:00 pm
Mathematics and Science Center: W201

Mathematics and Computer Science
Emory University