NSF PI Meeting on Science of Cloud Computing Application

PI information

Li Xiong
Assistant Professor
Department of Mathematics and Computer Science
Emory University
lxiong@mathcs.emory.edu

Research interests in Cloud Computing

Cloud Security, Privacy, and Auditing
Cloud Architecture and Systems

Current research activities related to Cloud Computing

The PI directs the Assured Information Management and Sharing (AIMS) research group at Emory University (http://www.mathcs.emory.edu/aims). Her general research goal is to develop models and techniques to enhance privacy, confidentiality, trust, interoperability, and scalability of distributed information systems. Her current research activities related to Cloud Computing include:

- Data de-identification and privacy preserving statistical data publishing. Privacy preserving data publishing releases a "sanitized" view of the sensitive data and provides a promising approach for sharing information while preserving data privacy. We have developed data de-identification and anonymization techniques with weak privacy notions mainly in the context of health information and multiple data providers in our previous research. Our current research tackles an open problem of non-interactive data release with differential privacy, a strong unconditional privacy, through a *data-driven* and *adaptive* querying framework. Our approach accesses the original database *indirectly* via an interactive mechanism that guarantees differential privacy and generates a differentially private view of the original data using a designed query strategy. We circumvent the hardness of the problem by novel and sophisticated use of the interface exploiting the characteristics of the underlying data. This research will be a promising building block for privacy preserving data outsourcing in the cloud.
- Scalable and extensible architecture and middleware for data federations. We have developed DObjects, a next-generation distributed mediator-based architecture for data federations, with dynamic query processing techniques. It consists of multiple decentralized system nodes which can serve as a mediator and wrapper or a mediator and form a *virtual system* in a P2P fashion. We are currently extending the middleware with secure distributed query processing services using secure multiparty computation techniques for securely aggregating data from multiple data sources. The architecture is flexible and extensible in that the system nodes form collaborative groups when necessary and can be easily deployed in the cloud while utilizing the elastic resources in the cloud. As an analogy, our system nodes can be considered as *droplets*, small elements that provide similar functionality in the cloud. Groups of *droplets* form a *micro-cloud* and our data federation framework can be considered as such *micro-cloud*.

Research Abstract

While Cloud Computing presents promising on-demand computing for cloud users to outsource their data and computations to the enormous data centers offered by cloud providers, data privacy and security is widely recognized as a major barrier for widespread adoption of Cloud Computing. Users are reluctant to place their sensitive data in the cloud with concerns about data disclosure to potentially untrusted cloud providers and other malicious parties. Part of the data confidentiality issues in the cloud reflects the well-established security challenges in the traditional data outsourcing setting. While homomorphic encryption schemes and non-encryption schemes such as data fragmentation have been studied, it remains an open challenge to support versatile and efficient computation on the outsourced data with assurances of confidentiality and privacy. In addition, data outsourcing in the cloud brings about new challenges and opportunities. What differentiates cloud-hosting providers from traditional hosting providers is their ability to offer *elastic* resources, purchasable in small time units at prices made possible through economies of scale. This challenges us to reexamine the existing solutions and design next generation secure data outsourcing techniques for the cloud that can systematically balance confidentiality and privacy requirements with computation needs of the data, and allow users to rapidly and dynamically provision their resource needs in order to realize the full potential of the elasticity and pay-per-use of the cloud.

If funded by NSF, the PI would like to fundamentally advance the research on data outsourcing in the cloud by developing an *integrated* and *adaptive* framework for secure and elastic data outsourcing to multiple cloud providers in the cloud. The specific research directions include:

- Integrated techniques that uniquely combine data encryption, partitioning, and statistical data outsourcing to ensure data confidentiality and privacy while minimizing the computation cost for both data preprocessing and query processing in the cloud. Each cloud provider may store parts of the data in original, encrypted, or statistical form that achieves differential privacy. A novel aspect of the research is to outsource statistical data with provable inference guarantee to the cloud server in addition to the data which may be partitioned or encrypted. Given a set of confidentiality and privacy constraints and an estimated workload, our problem is to design an optimal outsourcing arrangement for the cloud user consisting of proper encryption, fragmentation, and statics outsourcing that minimizes the cost associated with data preprocessing and the given query workload. Algorithms and heuristics will be developed to allow cloud users to systematically balance the confidentiality and privacy with the workload requirements.

- Adaptive techniques that allow cloud users to dynamically estimate the changing workload and data updates and rapidly adjust their outsourcing strategy to fully realize the potential of the elasticity and pay-per-use of the cloud. Controller mechanisms such as a simple Proportional-Integral-Derivative (PID) controller can be used to model and estimate current workload considering both its history and sudden changes. In addition to dynamic query workload, dynamic data presents important challenge and security implications. In particular, when a cloud user outsources multiple versions of data to the cloud, the different versions of partitions and statistical data will create additional disclosure risks from inference. Algorithms and heuristics will be developed to balance the data freshness, data accuracy, as well as outsourcing and computation cost while guaranteeing provable security.

The new techniques will provide a holistic conceptual foundation for secure data outsourcing in the cloud. In addition, a prototype system will be developed and evaluated through real healthcare applications using cloud platforms in collaboration with Dr. Vaidy Sunderam in the PI's department and the domain scientists at the Center for Comprehensive Informatics at Emory. Dr. Sunderam directs the Distributed Computing Lab (http://dcl.mathcs.emory.edu) and has considerable experience with resource sharing and heterogeneous computing systems with recent research focused on enhancing accessibility and portability across multiple cloud platforms. The systems techniques that make different types of cloud and local platforms compatible, host practical manifestations of remote databases, and perform at optimal levels will make the technology eminently usable. The domain scientists will provide real data and workload for evaluations to ensure the real impact of the research.