Name: K. Selçuk Candan

Affiliation: Arizona State University;
School of Computing, Informatics, and Decision System EnGineering;
Faculty of Computer Science and Engineering
Email: candan@asu.edu
Topics:

- Cloud Architectures and Systems
- Cloud Security, Privacy, and Auditing

Current Work:

- RanKloud (supported by HP Labs): Multimedia data is being produced in massive quantities and this data flood brings forth a need for highly parallelizable frameworks for scalable processing and efficient analysis of large media collections. We argue that when dealing with multimedia analysis, we need to consider utility as an inherent property of the data and features. When the utility is not uniform across the data and features, avoiding waste in processing necessitates task partitioning and resource allocation strategies that can prune unpromising objects from consideration without having to use available resources to enumerate results that will be eventually eliminated. RanKloud is an efficient and scalable utility-aware parallel processing system for analysis of large media data sets. Building on the MapReduce paradigm, RanKloud provides (a) adaptable, utility- and rank-aware data processing primitives (such as top-k, nearest neighbor, and skylines); (b) waste-avoidance strategies for utility-aware data partitioning and resource allocation; and (c) strategies for adaptation of the media processing workflows based on the utility-characteristics discovered in run-time. Results show that RanKloud provides significant gains by allocating resources in ways that consider the ranked semantics of the media analysis operations.
- Policy-Aware Data Processing in the Cloud (supported by HP Labs): we focus on challenges in protecting privacy of large scale healthcare databases. We consider a private data cloud based healthcare data management system, which provides a flexible (map-) based data model (as in Bigtable and HBase) that can potentially support complex data processing semantics under flexible data schema. In the context of policy-aware data clouds, not only queries on the data, but also the policy-management load (including audit log maintenance, policy-enforcement, and policy violation analysis) have to be considered. Since policy enforcement and on-demand policy verification and violation detection may require access to entire rows or entire columns row- or column-partitioning may have significant overheads if policy verification requires pulling together data and audit logs from different servers for each enforcement or analysis task. We are developing policy-aware partitioning strategies that ensure that (a) given knowledge about frequency of verification and enforcement of policies, the policy-management load on the system is minimized and the disclosure risks and audit-trail management load in the system is minimized.

Future Work: In 2009, President Obama stated that every American will have an electronic health record by 2014 and the stimulus bill allocated $36 billion to build electronic healthcare record systems. Such healthcare data management systems, however, impose stringent privacy and security requirements. In particular, in healthcare data management, the privacy of the individual information is protected by the Federal law. The Health Insurance Portability and Accountability Act (HIPAA) mandates regulations protecting the confidentiality of health information. Consequently, institutions dealing with healthcare data need to have established mechanisms, policies, and procedures for annotating records to show in what ways information contained in them has been accessed or disclosed. Audits often require specific physical location of a data entry (on a server or table) need to be known and that the data entry has not been accessed or modified without a record and has never been exchanged (within or across systems) in an insecure manner. The audit analysis mechanism needs to be flexible to detect violations of any stated policy as well as any patterns of attack in an on-demand manner.

Cloud computing platforms can provide high degrees of data processing parallelism to provide scalability of large scale data analysis applications. On-demand policy analysis, therefore, can benefit from large scale and highly parallelizable data analysis frameworks, such as RanKloud. This, however, requires mappings from policy specifications (e.g. in WS-DataAssurancePolicy) and corresponding workflows to data processing workflows that can identify violations of a stated policy within a large health-care data cloud. Therefore we need to tackle scalable on-demand policy verification over audit trails through scalable data processing frameworks. In particular, we plan to consider mappings from policy specifications to data processing workflows for on-demand policy verification and attack detection.

- Challenge I: Since on-demand policy verification and violation detection may require access to entire rows or entire columns; how can we partition the data? Row- or column-partitioning may have significant overheads if policy verification requires pulling together data and audit logs from different servers for each request. Can different replicas may use different partitioning strategies so that different classes of policy structures are efficiently managed?
- Challenge II: Development of a distributed audit management protocol suitable for scalable, large scale data management. Development of a secure, distributed indexing and search mechanism for locating relevant data and/or audit log entries within a data cloud for on-demand policy analysis.
- Challenge III: Scalable on-demand policy verification over audit trails through scalable data processing frameworks. Mapping from policy specifications to data processing workflows for on-demand policy verification or attack detection. Investigation of self-reporting data processing frameworks, which can create audit-trails for derived, integrated data they create during data processing.
- Challenge IV: Design of a scalable on-demand policy verification and audit analysis system, building on our current utility-aware scalable data analysis framework, RankLoud that is being designed to support massive data processing and decision making applications, where (a) precision of the retained data elements and operations on these data are variable and (b) the data matching process is inherently imprecise.