Towards a General Framework for Secure MapReduce Computation on Hybrid Clouds

Chunwang Zhang, Ee-Chien Chang, Roland H.C. Yap
School of Computing, National University of Singapore
{chunwang, changec, ryap}@comp.nus.edu.sg

Keywords
Data security; MapReduce; hybrid clouds; information leakage

The idea of a *hybrid cloud* is to combine a private cloud (e.g., an organization's in-house private datacenter) together with a public cloud (e.g., Amazon EC2). Hybrid cloud computing offers increased scalability and cost-effectiveness: the private cloud can be used for typical workloads, but when additional resources are needed during peak computations, the public cloud is harnessed. This hybrid cloud architecture has already gained adoption [1] and is still undergoing rapid development [4].

However, hybrid cloud computing needs to address the confidentiality and privacy issues on public clouds. Security and privacy are ranked as the top concerns for organizations considering moving their applications and data to the cloud [2, 4]. There are good reasons for these concerns, e.g., Ristenpart et al. [8] demonstrate that confidential information can be extracted through side-channel information leakage in VMs. Many data breaches have been reported for various cloud service providers [3,5]. On the other hand, organization data often involve both sensitive and non-sensitive information. For example, an organization's filesystem may contain both general files mixed with confidential business data. Also, many datasets for analytical tasks such as network logs and healthcare records may involve data from public sources with private organization data. Computations on such mixed-sensitivity data should not be carried out on the public cloud without protection to prevent data leakages. Cryptographic techniques such as fully homomorphic encryption [7] that enable computation on encrypted data are still far from efficient for large data.

With a hybrid cloud, one solution may be to separate the computation on non-sensitive data from that on sensitive data, such that the former can be comfortably outsourced to the public cloud while the latter, possibly much smaller in size, can be easily handled on the private cloud. In this way, the computation can be carried out both securely and efficiently. However, this hybrid computing model is not supported by today's data-intensive computing frameworks. In particular, MapReduce [6] (MR) is designed for only one cloud and does not distinguish between data and servers with differing sensitivities. A cloud user who wants to run MR jobs with mixed-sensitivity data on a hybrid cloud needs to manually split the data, compute each partition on the corresponding cloud independently and combine the results in her own code. What is desired is an automatic and general framework to facilitate secure computing on hybrid clouds and we focus on MR. Sedic [9] addresses this problem to some degree but has limitations in terms of flexibility and support for complex MR jobs.

We work towards this direction by providing a general hybrid MR framework which deals automatically with mixed sensitivity MR jobs while supporting new kinds of MR programming which manipulate the data sensitivity. We propose a general tagged-MapReduce that (conceptually) augments each key-value pair in MR with a sensitivity tag and extends the map and reduce functions appropriately which: 1) allows fine-grained dataflow control during execution and supports scheduling of map and reduce tasks in the two clouds; 2) allows programmers to code sophisticated policies to guide sensitivity transformation during execution; 3) provides sensitivity information for data across multiple MR jobs necessary for complex MR computations with chained jobs. Sedic programs are a special case of our model but Sedic cannot express all tagged-MR programs. However, a hybrid MR framework may sacrifice in performance due to the security constraint. Our tagged-MR has a security model and scheduling mechanisms to deal with these issues.

SoCC'13, Oct 01–03 2013, Santa Clara, CA, USA
ACM 978-1-4503-2428-1/13/10.
http://dx.doi.org/10.1145/2523616.2525944

References

[1] Forecast for 2010: The Rise of Hybrid Clouds. Online at `http://gigaom.com/2010/01/01/on-the-rise-of-hybrid-clouds/`, 2010.

[2] AMD 2011 Global Cloud Computing Adoption, Attitudes and Approaches Study. Online at `http://www.slideshare.net/AMD/amd-cloud-adoption-approaches-and-attitudes-research-report`, 2011.

[3] Epsilon Data Breach Highlights Cloud Computing Security Concerns. `http://www.eweek.com/c/a/Security/Epsilon-Data-Breach-Highlights-Cloud-Computing-Security-Concerns-637161/`, 2011.

[4] 2012 Cloud Computing Survey. Online at `http://northbridge.com/2012-cloud-computing-survey`, 2012.

[5] Dropbox: Yes, we were hacked. Online at `http://gigaom.com/2012/08/01/dropbox-yes-we-were-hacked/`, 2012.

[6] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating System Design and Implementation*, pages 137–150, 2004.

[7] C. Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 169–178, 2009.

[8] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pages 199–212. ACM, 2009.

[9] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacy-aware data intensive computing on hybrid clouds. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, pages 515–526. ACM, 2011.