CS554 Project Ideas

FusionFS:Prov – Scalable Distributed Data Provenance

Overview

It has become increasingly important to capture and understand the origins and derivation of data (its provenance). A key issue in evaluating the feasibility of data provenance is its performance, overheads, and scalability. In this project, we will design and implement a distributed provenance system at the file system level. Particularly, you will extend the current system [1] with new architectures (e.g. switching from hash tables to binary sorted trees) for storing the provenance information, and implement a programmatic interface for provenance query. This provenance system will be patched to FusionFS in the next release.

Relevant Systems and Reading Material

Please read the following papers (and their references) before submitting your proposal:

[1] Dongfang Zhao, Chen Shou, Tanu Malik and Ioan Raicu. Distributed Data Provenance for Large-Scale Data-Intensive Computing, *IEEE International Conference on Cluster Computing*, 2013. Available online: http://datasys.cs.iit.edu/~dongfang/download/pafs_crc.pdf

[2] Chen Shou, Dongfang Zhao, Tanu Malik and Ioan Raicu. Towards a provenance-aware distributed filesystem, *5th USENIX Workshop on the Theory and Practice of Provenance*, 2013. Available online: http://datasys.cs.iit.edu/~dongfang/download/pafs.pdf

Preferred/Required Skills

Principles: operating system, distributed systems, computer network, database systems

Programming: Shell Script, Perl/Python, C, C++, PThread, sockets, FUSE

Operating systems: Linux

Project Mentor

Dongfang Zhao

Email: dzhao8@hawk.iit.edu