CAREER: Lawful Storage and Processing of Personal Data

Supreeth Shastri, University of Iowa

Overview

Our society is in the midst of granting a new right to people: *privacy and protection of personal data*. A prominent example is the General Data Protection Regulation (GDPR), introduced in Europe in 2018, which has since then emerged as a model regulation for similar efforts in the U.S. and around the world. GDPR requires new types of control- and data-plane operations from systems that store and process personal data. Yet, widely used data analytics platforms such as Hadoop offer little or no support for GDPR, instead leaving the burden solely to applications. This is causing significant challenges in the computing and data science communities, which is evident in the fact that GDPR fines are being issued, on average, once every 1.4 days.

The overall objective of this proposal is to build system software to enable lawful storage and processing of personal data. The proposal is rooted in our insight that lawful computing requirements of GDPR can be distilled into a layered systems stack (that we call, *GDPR Stack*). We will build system software to make the Hadoop ecosystem GDPR capable via three objectives: (i) lawful storage of personal data with *GDPR-aware HDFS*, (ii) lawful computing on personal data with *GDPR-aware MapReduce*, and (iii) enabling applications to manage end-to-end compliance with *GDPR orchestrator for Hadoop*. If successful, this project would enable the broader computing community of data scientists, application programmers, and system administrators to design, develop and deploy lawful personal-data applications, easily and efficiently.

We also recognize how the CS education has currently relegated lawful and ethical data processing as optional topics to be taught in niche courses. We propose an educational plan that includes upgrading the CS curriculum, writing a textbook, and creating teaching materials to train the next generation to be thoughtful technologists. In summary, this proposal lays the foundation for the PI's long-term goal of creating scientific methods and artifacts to help the computing community comply with data regulations and for people to exercise their data rights.

Intellectual Merit

This project will advance the state of knowledge in computing- and database systems research regarding: (i) how to distill the lawful computing requirements of data regulations into a layered systems stack (using GDPR as the reference law); (ii) how to systematically introduce GDPR capability into legacy data analytics platforms (using Hadoop as the working example); and finally, (iii) how to design APIs, mechanisms, and policy controls at the systems level so that application developers can achieve GDPR compliance, easily and efficiently. Thus, we are confident the project will make a foundational contribution to the emerging interdisciplinary area of *lawful and ethical data science*.

Broader Impacts

This proposal address an important socio-technical problem: the rapid proliferation in the use of personal data and the subsequent emergence of data regulations has made building lawful personal-data applications, *a problem of the many rather than a problem of a few*. The research aims of this project, centered on the widely used Hadoop ecosystem, will benefit the broad data science community. The educational aims of this project, centered on the theme of training thoughtful technologists, will foster the development of a globally competitive STEM workforce. Ideas and artifacts resulting from this project will offer the PI a unique opportunity to increase the partnership between academia, industry, and others including policy makers, enforcement agencies, and data protection officers.

Keywords: Data regulations, Lawful computing, GDPR, Hadoop, HDFS, MapReduce