MIT and Harvard University researchers have developed a platform, called Riverbed, that ensures that web services adhere to users' preferences on how their data are stored and shared in the cloud.

Image: Chelsea Turner, MIT

Putting data privacy in the hands of users

New platform acts as a gatekeeper to ensure web services adhere to a user's custom data restrictions.

Rob Matheson | MIT News Office
February 20, 2019

A new platform developed by MIT and Harvard University researchers ensures that web services adhere to users' preferences on how their data are stored and shared in the cloud.

In today's world of cloud computing, users of mobile apps and web services store personal data on remote data center servers. These data may include photos, social media profiles, email addresses, and even fitness data from wearable devices. Services often aggregate multiple users' data across servers to gain insights on, say, consumer shopping patterns to help recommend new items to specific users, or may share data with advertisers. Traditionally, however, users haven't had the power to restrict how their data are processed and shared.

In a paper being presented at this week's USENIX Networked Systems Design and Implementation conference, the researchers describe a platform, called Riverbed, that forces data center servers to only use data in ways that users explicitly approve.

In Riverbed, a user's web browser or smartphone app does not communicate with the cloud directly. Instead, a Riverbed proxy runs on a user's device to mediate communication. When the service tries to upload user data to a remote service, the proxy tags the data with a set of permissible uses for their data, called a "policy."

Users can select any number of predefined restrictions — such as, "do not store my data on persistent storage" or "my data may only be shared with the external service x.com." The proxy tags all the data with the selected policy.

In the datacenter, Riverbed assigns the uploaded data to an isolated cluster of software components, with each cluster processing only data tagged with the same policies. For example, one cluster may contain data that can't be shared with other services, while another may hold data that can't be written to disk. Riverbed monitors the server-side code to ensure it adheres to a user's policies. If it doesn't, Riverbed terminates the service.

Paper: "Riverbed: Enforcing User-defined Privacy Constraints in Distributed Web Services."

Frank Wang

Computer Science and Artificial Intelligence Laboratory

Department of Electrical Engineering and Computer Science

School of Engineering

The privacy risks of compiling mobility data

Private browsing gets more private

Protecting web users' privacy

Riverbed aims to enforce user data preferences, while maintaining advantages of cloud computing, such as performing large-scale computations on outsourced servers. “Users give a lot of data to web apps for services, but lose control of how the data is used or where it’s going,” says first author Frank Wang SM ’16, PhD ’18, a recent graduate of the Department of Electrical Engineering and Computer Science and the Computer Science and Artificial Intelligence Laboratory. “We give users control to tell web apps, ‘This is exactly how you can use my data.’”

On that thread, an additional perk for app developers, Wang adds, is establishing more trust with users. “That’s a big thing now,” Wang says. “A selling point for your app would be saying, ‘My app’s goal is to protect user data.’”

Joining Wang on the paper are PhD student Ronny Ko and associate professor of computer science James Mickens, both of Harvard.

Creating “universes”

In 2016, the European Union passed the General Data Protection Regulation (GDPR), which states that users must consent to their data being accessed, that they have the right to request their data be deleted, and that companies must implement appropriate security measures. For web developers, however, these laws provide little technical guidance for writing sophisticated apps that need to leverage user data.

In the past, computer scientists have designed “information flow control” (IFC) systems that allow programmers to label program variables with data policies. But with so many variables and many possible interactions between variables, these systems are difficult to program. Thus, no large-scale web services use IFC techniques.

Primarily, Riverbed leverages the fact that the server-side code of an app can run atop a special “monitor” program — programs that track, regulate, and verify how other programs manipulate data. The monitor creates a separate copy of the app’s code for each unique policy assigned to data. Each copy is called a “universe.” The monitor ensures that users who share the same policy have their data uploaded to, and manipulated by, the same universe. This method enables the monitor to terminate a universe’s code, if that code attempts to violate the universe’s data policy.

This process incorporates a custom interpreter, a program that compiles programming language into code that’s understood by a computer. Interpreters are also used to help runtime programs implement low-level commands into an original program as it runs. The researchers modified a traditional interpreter to extract defined policies from incoming user data and labels certain variables with specific policy direction. Labels will, for instance, denote whitelisted web services for data sharing or restrict persistent storage — meaning the data can’t be stored when the user stops using the web service.

“Say I want my data to be aggregated with other users. That data is put into its own universe with other user data with the same policy,” Wang says. “If a user doesn’t want to share any data with anyone, then that user has their own whole universe. This way, you don’t have any cross-pollination of data.”

For developers, this makes it much easier to comply with GDPR and other privacy laws, Wang says, because users have given explicit consent for data access. “All users in each universe have the same policies, so you can do all your operations and not worry about what data is put into an algorithm, because everyone has the same policy on data in that universe,” Wang says.

Efficient copying

In the worst-case scenario, Wang says, each user of each service would have a separate universe. Generally, this could cause significant computation overhead and slow down the service. But the researchers leveraged a relatively new technique, called “container-based virtualization,” which allow the Riverbed monitor to more efficiently create multiple universes of the same program. As a result, universe management is fast, even if a service has hundreds or thousands of universes.

In their paper, the researchers’ evaluated Riverbed on several apps, demonstrating the platform keeps data secure with little overhead. Results show that more than 1,000 universes can squeeze onto a single server, with added computation that slows down the service by about 10 percent. That’s fast and efficient enough for real-world use, Wang says.

The researchers envision the policies as being written by advocacy groups, such as Electronic Frontier Foundation (EFF), an international nonprofit digital rights group.

New policies can be “dropped in” to a Riverbed-run service at any time, meaning developers don’t need to rewrite code.

Topics: Research | Computer science and technology | Privacy | Mobile devices | Data | Technology and society | Cyber security | Computer Science and Artificial Intelligence Laboratory (CSAIL) | Electrical Engineering & Computer Science (eecs) | School of Engineering | Alumni/ae