TECHNISCHE UNIVERSITÄT BERLIN

Bachelor's/Master's thesis

Discovery and Inventory of Personal Data in Distributed Service Environments

Context

The EU General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) define a strong regulatory framework for data protection/privacy by design and by default. At the same time, more and more services do collect personal data from countless data subjects – especially the well-known big players. However, also smaller enterprises face the challenges of being compliant to all regulations (while fearing severe fines).

State of the Art & Problem

Organizations are required to establish technical and organizational measures as safeguards against potential data breaches. At the same time, supervisory authorities are expanding their activities to control data controllers (and processors). In this context, distributed systems (that may follow microservice architectures, multi-cloud strategies etc.) are especially hard to "x-ray" for personal data. First solution approaches from industry become visible [1], however, they lack algorithmic transparency and transferability to non-proprietary use-cases. Moreover, the clear definition of personal data is subject to current research (e.g. [2]).

Thesis Topic & Goal

In this thesis, a method to discover personal data in selected paradigmatically different storage and processing settings (Relational DBs, NoSQL stores, Object stores, Stream processing systems etc.) is to be conceptualized and practically implemented. Doing this, effective means need to be found or used, that identify common patterns or context-specific indicators of personal data. In order to create a realistic testbed containing appropriate data sets, a representative distributed cloud application is to be set up [3] [4]. Secondly, an integrating inventory of all locations of personal data is to be developed. Said inventory should be able to reflect the common scenario of replicated data stores across data centers and/or potential data integration strategies. Meanwhile, changes in service compositions occur regularly and the inventory needs to keep track of these. This approach can be inspired by and compared to existing purpose-based access control mechanisms, transparency information representation and transparency-related service descriptions. Eventually, the method should be designed to improve the internal and external auditability (detection of risks of data breaches) [5]. Technical scope (incl. choice of cloud services), approach (tracing, logging, monitoring…), etc. are to be specified further together with the candidate based on his/her skills, experiences and interests.

Contact: Elias Grünewald
eg@ise.tu-berlin.de

Skills

- Good (web) programming skills
- Good knowledge about distributed systems and different data management systems
- Interest in the development of privacy enhancing technologies

References

[1] https://aws.amazon.com/de/macie
[2] Finck, M., & Pallas, F. (2020). They who must not be identified-distinguishing personal from non-personal data under the GDPR. International Data Privacy Law, pp. 19-14.
[3] Cf. https://github.com/GoogleCloudPlatform/microservices-demo or https://github.com/didier-durand/microservices-on-cloud-kubernetes
[4] If real data sets are not available, see: https://github.com/lk-geimfari/mimesis or https://github.com/danibram/mocker-data-generator
[5] Silva, P., Gonçalves, C., Godinho, C., Antunes, N., & Curado, M. (2020, July). *Using NLP and Machine Learning to Detect Data Privacy Violations*. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 972-977. IEEE.

Our Mission:

Our lectures cover fundamental methods and techniques in the areas of service computing, cloud computing, and enterprise computing. We like to engage students in hands-on building of distributed information systems and to take an interdisciplinary approach to evaluating such systems. Through a close mentoring of students, especially in our seminars, we aim to introduce students to our ongoing research and to excite them to do future studies and research with us.

ISEngineering
Information Systems Engineering