# In-Storage Data Transformations for Enforceable Privacy

Claudiu Mihali (presenter, student)
Anca Hangan Gheorghe Sebestyén
Technical University of Cluj-Napoca, Romania

Zsolt István
IMDEA Software Institute, Madrid, Spain

## 1. Motivation and Problem Statement

Big Data has revolutionized our lives but has also created challenges in the areas of privacy and data protection. As a result, there is a push for stronger data protection rules and companies are investing in tools for tighter control and monitoring of personally identifiable information (PII) processed and stored in their systems. Enforcing privacy-related rules within large processing pipelines, however, is still very much an open challenge because computation is often split across a large number of nodes. We focus on the question of how to always comply with regulations at the data source level without negatively impacting performance, while, at the same time, avoiding being overly restrictive in data use? A solution can come from combining the state of the art in security and privacy with high bandwidth in-network and near-data processing: Using specialized hardware devices to implement a system capable of modifying the data that comes from a storage system on the fly, offering a version with the same schema but that is stripped from its sensitive contents while still being able to be useful in initial exploration for data analytics pipelines and machine learning models that will eventually consume it.

## 2. Our Prototype

Data perturbation is a privacy preservation technique that alters the values of elements in a database in order to disguise the sensitive information while preserving the particular data properties that are critical for building meaningful data analytics models. This technique can often be reduced to simple operations applied to rows, columns or individual records of data, which makes it suitable for the implementation with hardware devices. The main challenge is to select a transformation that balances the privacy protection of the new data with the remaining utility, which are normally considered as a pair of conflicting factors.

Differential privacy (DP) is another relevant tool that provides a strong privacy guarantee. However, its main limitation is the privacy budget constraint which involves implementing a mechanism that allows only a limited number of repeated queries to be performed before some information leakage occurs. The data perturbation method targets different data processing operators than DP and does not require keeping track of such budgets – nonetheless, our prototype platform allows future integration of DP solutions.

The work presented in the poster builds on an FPGA-based Key-value Store [1] that offers capabilities for storing and accessing Parquet files. In order to be able to implement transparent data perturbations at line rate (10Gbps), we have implemented a module for data decompression, different accessing patterns (row-based, column-based, individual records) and are working on modules for data transformation. These near-data compute modules can be activated or deactivated as a decision of users with high access privileges.

The work flow from a user's perspective is to (1) store the compressed data in the KVS as Parquet files with relevant metadata; (2) access selectively row batches of certain columns from the stored files which are being transparently decompressed and transformed, ready to be fed to the data analytics pipelines; (3) after the initial exploration/prototyping phase is over for the ML models or analytic pipelines, higher privilege users can disable the transformation modules for accurate results.

## 3. Ongoing and Future Work

Currently we have implemented a variant of geometric perturbation with 3D Rotation [2]. With this method we precompute a rotation matrix which is stored in the KVS alongside the Parquet file metadata. The transformation on retrieval will consist of grouping values from rows in triplets and rotating these triplets with the rotation matrix. Several data mining models such as kNN classifiers, kernel methods, SVM classifiers and linear classifiers are invariant to the rotation transformation [3].

Our goal with the poster is to showcase our initial results and to receive feedback, on the one hand, about other possible transformations to perform and, on the other hand, the possible challenges one could face when deploying such "smart storage" nodes in production systems.

## References

[1] L. Kuhring, E. Garcia, Z. Istvan. Specialize in Moderation – Building Application-aware Storage Services using FPGAs in the Datacenter. In *HotStorage'19*.

[2] S. Upadhyay, C. Sharma, and others. Privacy preserving data mining with 3-D rotation transformation. In *Journal of King Saud University – Computer and Information Sciences 2016*.

[3] K. Chen, L. Liu. Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining. In *Knowledge and Information Systems'11*.