# In-Storage Data Transformations for Enforceable Privacy

**Claudiu Mihali**, Anca Hangan, Gheorghe Sebestyén (TUCN), Zsolt István (IMDEA)

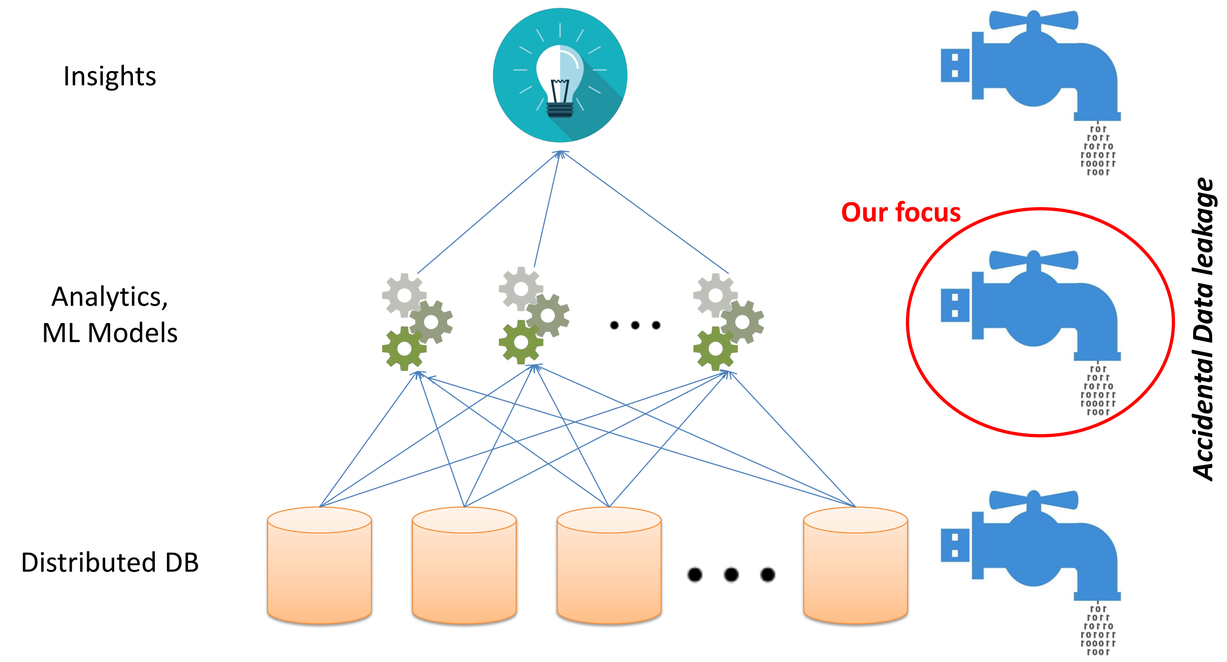
## Motivation

### Growing Data Sizes → Distributed Systems → More difficult to protect

• Companies **store** and **process** an increasing amount of data, some of which is **sensitive** or **identifiable**

• Data processing pipelines are becoming more complex: **larger**, **distributed across multiple nodes**

• **Consequence**: It is becoming more and more difficult to enforce privacy-related rules




## Contribution

### In-storage data perturbation

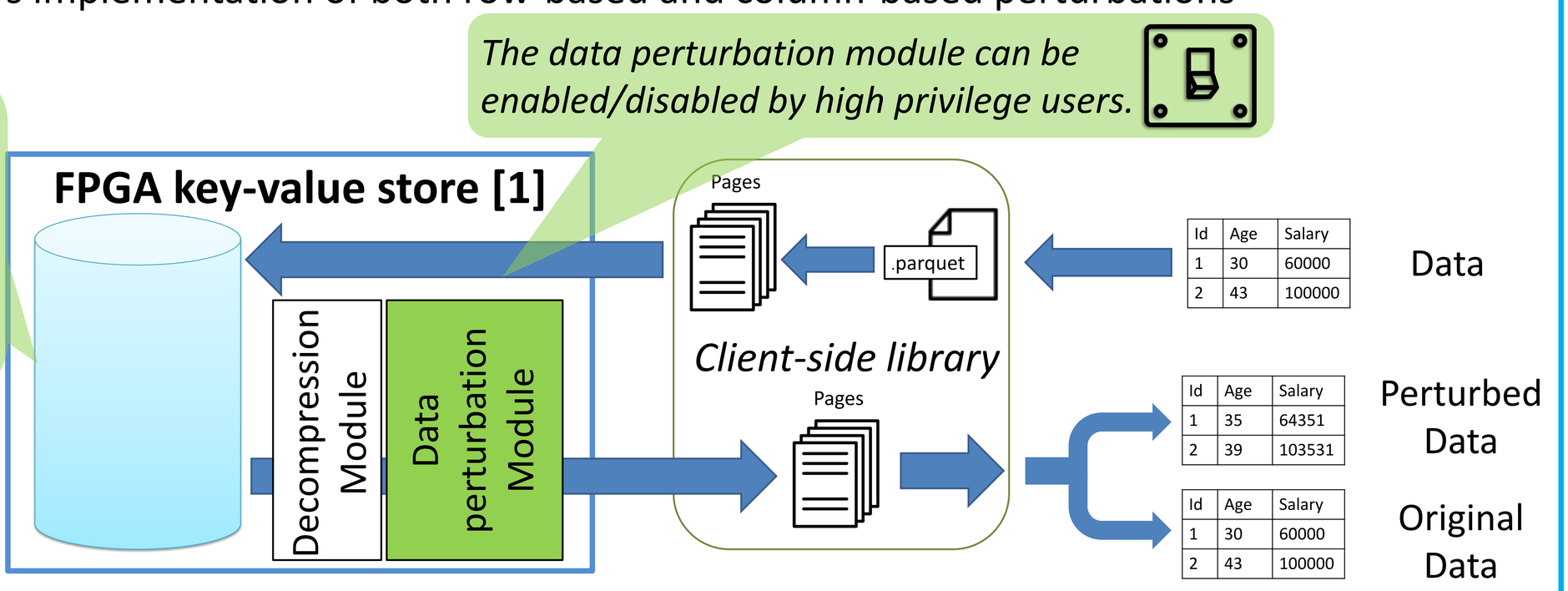
• Data perturbation = altering the values of elements in a database in order to disguise the sensitive information while preserving the particular data properties that are critical for building meaningful data analytics models

• It is cheap to perform and can often be reduced to some simple operation applied to rows, columns or individual records of data → suitable for implementation on FPGAs

• Our framework allows implementation of both row-based and column-based perturbations

*More **flexible** than differential privacy (repeated queries do not leak information), but not as strong of a privacy guarantee.*

*The data perturbation module can be enabled/disabled by high privilege users.*




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

## Applications

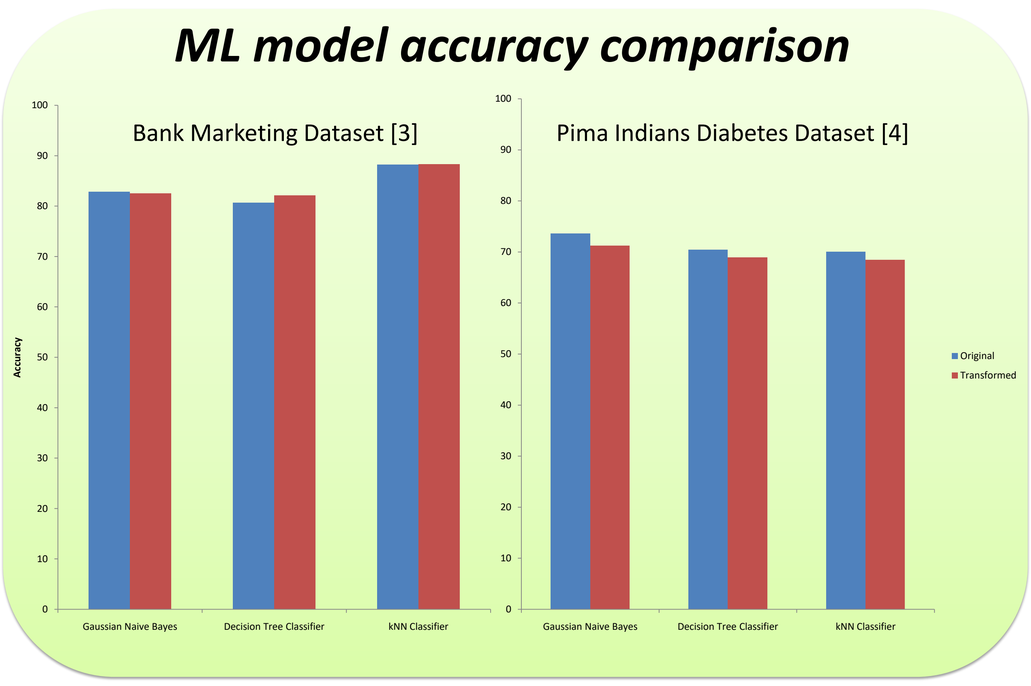
### Privacy preserving data analytics with 3D rotation transformation

• The 3D rotation transformation consists of randomly partitioning the set of columns into triplets and rotating each triplet with an orthogonal rotation matrix.

• The rotation matrix is chosen in order to maximize the variance of the difference between the original and perturbed data [2]. It will be stored in the KVS alongside the Parquet file metadata.

• The rotation transformation preserves the geometric properties that many data analytics models are based on.

***ML model accuracy comparison***

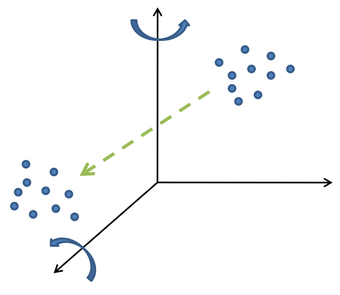



| Col 1 | Col 2 | ... | Col n |
|---|---|---|---|
| | | | |
| | | | |

Random partitioning

Set of 3D points:
$\{ (col_x, col_y, col_z) \mid (x, y, z) =$ 3-permutation of the column set$\}$

3D Rotation



[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] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.
[4] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.