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

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.

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.

Privacy preserving data mining with 3D rotation transformation. In Journal of King Saud University – Computer and Information Sciences 2016.

Applications

| Original | Perturbed
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
<td>Salary</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Salary</td>
<td>Salary</td>
</tr>
<tr>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Salary</td>
<td>Salary</td>
</tr>
</tbody>
</table>

3D Rotation

3-Perturbation ML model accuracy comparison


