upro EU
PROMOTIONAL PARTNERSHIP

In-Situ Anonymization of Big Data

Tomislav Križan
Consultancy Director @ Poslovna Inteligencija d.o.o

Abstract

- Vast amount of data is being generated from versatile sources and organizations are primarily interested in analyzing it for further information and improved decision making.
- Data analysis surely leverages data value for the organization and opening data publicly benefits wider community and fosters further scientific research.
- However, analysis means disclosing data to other employees, third party organizations or even publishing it on the internet for everybody to use.
- Stored data is often secret, proprietary or protected by data protection and privacy laws and organization can suffer significant financial loss if that data is misused.
- Huge cost of data theft and misuse urged organizations to improve security measures and protect data

Why is data privacy required?

- Production environment
  - security model to control access
- Non-production environment
  - security is opened up to enable development and testing
- Non-production business drivers
  - Development
  - Testing
  - Support
  - Outsourcing

Data Privacy

- Secure enables data privacy by providing robust data masking functionality.
- What is Data Masking?
  - Protecting sensitive information by hiding or altering data so that an original value is unknowable.
- Also known as:
  - De-identifying
  - Protecting
  - Camouflaging
  - Data masking
  - Data scrubbing

Why Data Masking?

- You use Data Masking when you need to share database tables with sensitive information.
- Data Masking creates a copy of the production database and lets you replace sensitive information with custom values.
- Benefits:
  - Hides sensitive information
  - Creates a copy of the production database
  - Allows you to create realistic test data

Legal Requirements

- EU Data Protection Directive (e.g. „Directive 95/46/EC of The European Parliament and of the Council“ adopted on October 24th 1995. and further elaboration through „Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques“ adopted on April 10th 2014.) **protects the privacy** of all personal data about citizens of the EU.

Legal Requirements

- It especially addresses processing, using and exchanging data. In this directive, personal data is defined as “any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

Legal Requirements

- *European Data Privacy Directive* regulates disclosing and processing of personal data and aims to be applied throughout the EU equally
- In U.S. there is no such single officially accepted directive, but there are various acts and industry standards:
  - *Payment Card Industry Data Security Standard (PCI-DSS)* specifies twelve technical requirements for the environment where cardholders' data is stored
  - *Health Insurance Portability and Accountability Act (HIPAA)* is a comprehensive act covering health service quality
  - *The Gramm-Leach-Bliley Act (GLBA)* protects financial data.

Sensitive data

- Sensitive data attributes can be classified into several types regarding their sensitivity:
  - *Direct (explicit) identifiers* contain values that can uniquely identify persons or organizations (i.e. sensitive entities). Examples are name, telephone number, email address, credit card number, various IDs, etc.
  - *Indirect (implicit, quasi) identifiers* can identify the entity but only indirectly. Examples include geographical location, gender, marital status, profession, date of birth, etc.

Data Anonymization System Requirements

- A system for anonymization should adhere to several requirements:
  - ***Security*** should be strong enough to make any tracing of masked (anonymized) values back to original sensitive information hard enough or impossible
  - ***Speed*** is measured in number of records masked in unit of time. This capability determine whether the system is capable of working online (real-time mode, data in motion) or only offline (batch mode, data at rest).
  - ***Techniques***. Supported algorithms should be able to anonymize/generate/format all the different value types of the dataset (credit card numbers, zip codes, telephone numbers, etc.).

Data Anonymization System Requirements

- ***Connectivity***. With data commonly being stored in databases, the system could offer connecting to various databases, creating queries and preserving physical structure, indices, primary keys, triggers and referential integrity across tables. For input from file, various file formats (fixed width records, delimiters and encodings) should be supported. Output can also be of different types.
- ***Resilience***. The system should skip/report invalid data such as improper or missing values in a record/row. Marking the last successful record (through use of checkpoints) would enable restarting the masking from before the record of last failure.

Data Anonymization Techniques

- There are two main approaches to anonymization based on:
  - ***Randomization*** - comprises any modification of the original values either by removing them (nullifying), replacing them with randomly generated or chosen values, adding noise or swapping.
  - ***Generalization*** - techniques convert values into more general ones

Data Anonymization Techniques (characteristics)

- Techniques can be further described with several characteristics:
  - *Deterministic (repeatable)* - same original value always maps into the same masked value.
  - *Uniqueness* – two different original values cannot map into the same masked value. Conditional sensitive - different masking for different given conditions/parameters.

Data Anonymization Techniques (characteristics)

- *Partial* – masking only part of the value (e.g. only several last digits of the telephone number).
- *Realistic* – masked output looks meaningful. Length preservation - masked value is of the same length as the original.
- *Reversibility* – it is possible to get original value from the masked one (e.g. provided the encryption key).
- *Information loss* – how much information is lost during the masking.

Data Anonymization Techniques (methods)

- *Substitution* technique replaces the original value with some other fake masked value
- *Shuffling or data swapping* randomly rearranges values inside one dataset column while preserving the order in other columns
- *Blurify* technique changes the original value by a given variance

Data Anonymization Techniques (methods)

- *Nulling out* simply deletes the original value
- *Character masking* technique is similar to nulling out and replaces the original value with a specified character constant
- *Encryption* technique encrypts the original value into anonymized one by using the provided key

Data Anonymization Techniques (methods)

- *Hashing* is a form of encryption technique where no key is used and data of arbitrary length is irreversibly masked into fixed length message digest output
- *Aggregation* aggregates several clustered records and replaces their values with generalizations

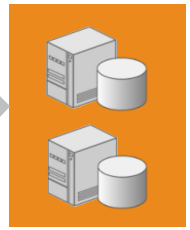
Data Anonymization Techniques (methods)

- *K-anonymity* is a requirement that pseudo-identifiers of each record must match at least *k* other records in the anonymized dataset. Optimal k-anonymization is an NP-hard problem and there are various approaches to meet this requirement. *L-diversity* and *t-closeness* models are refinements of *k-anonymity.*

Innovation Driver

- Growing volume of data and privacy concerns spawned work both in newer anonymization techniques and their merging with popular large scale distributed computing systems developing better anonymization techniques to be more resilient to background knowledge and linking attacks done by using increasingly available online datasets.

Data Masking Concept



- Development and Testing
- Training
- Support
- Data Analysis
- Outsourcing and Offshoring

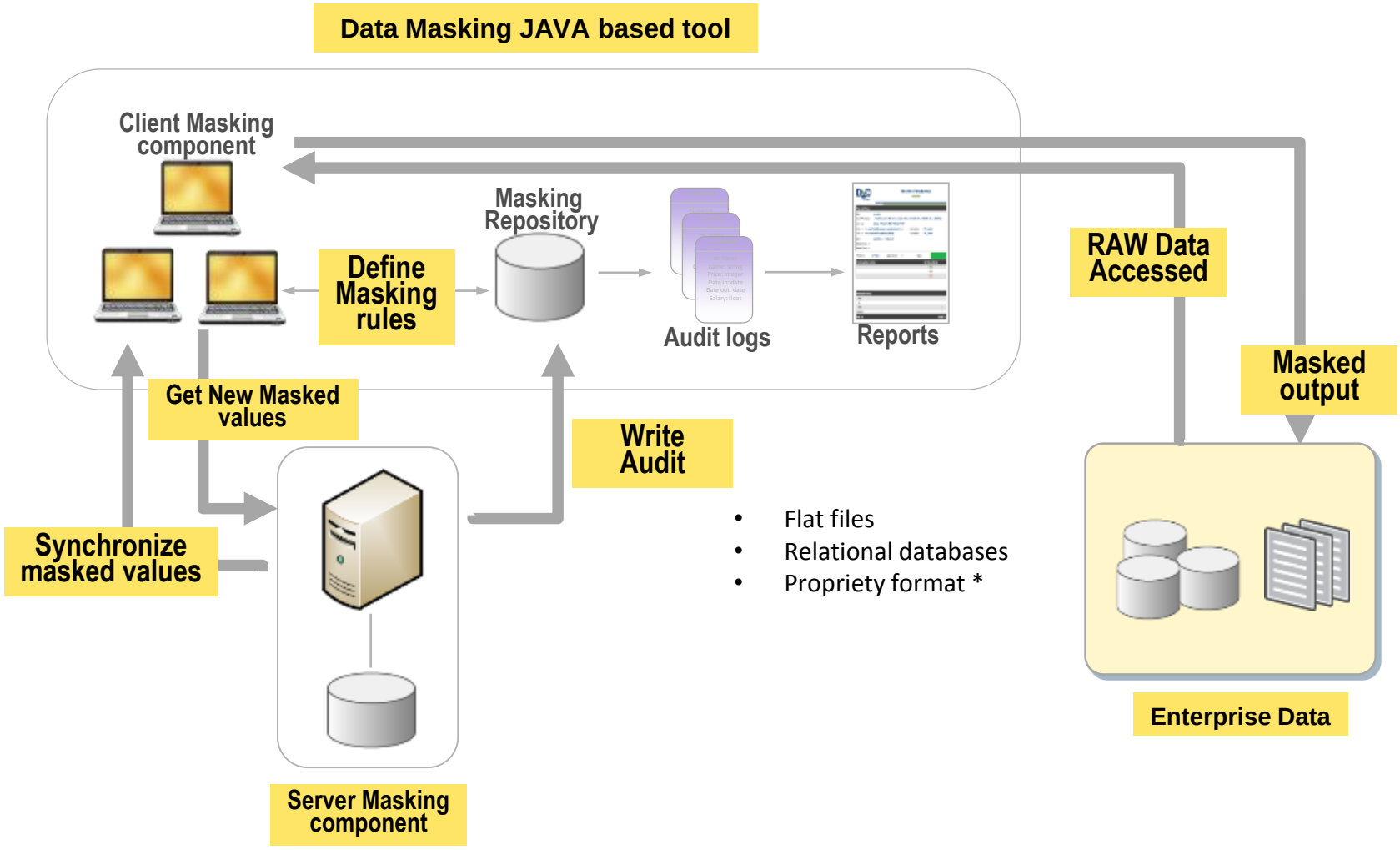
John Smith
654-65-8945
4563-3456-9876-6342
100 Cardinal way
Redwood city

Glen Carter
900-45-2643
4563-XXXX-XXXX-6342
342 54th Street
New York

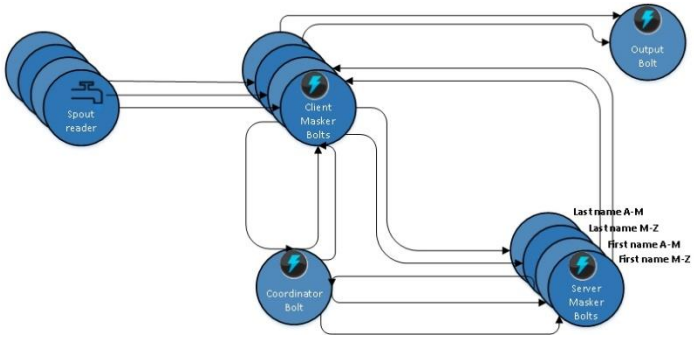
In-Situ Anonymization system

- The anonymization system in question was developed for the purposes of European FP7 FERARI (Flexible Event pRocessing for big dAta aRchItectures) project
- Logically, solution consists from **Client component** which reads and writes data, and **Server component** which provides algorithm and masked values to any number of Client components thus allowing parallel execution which will provide consistent result across different Client components

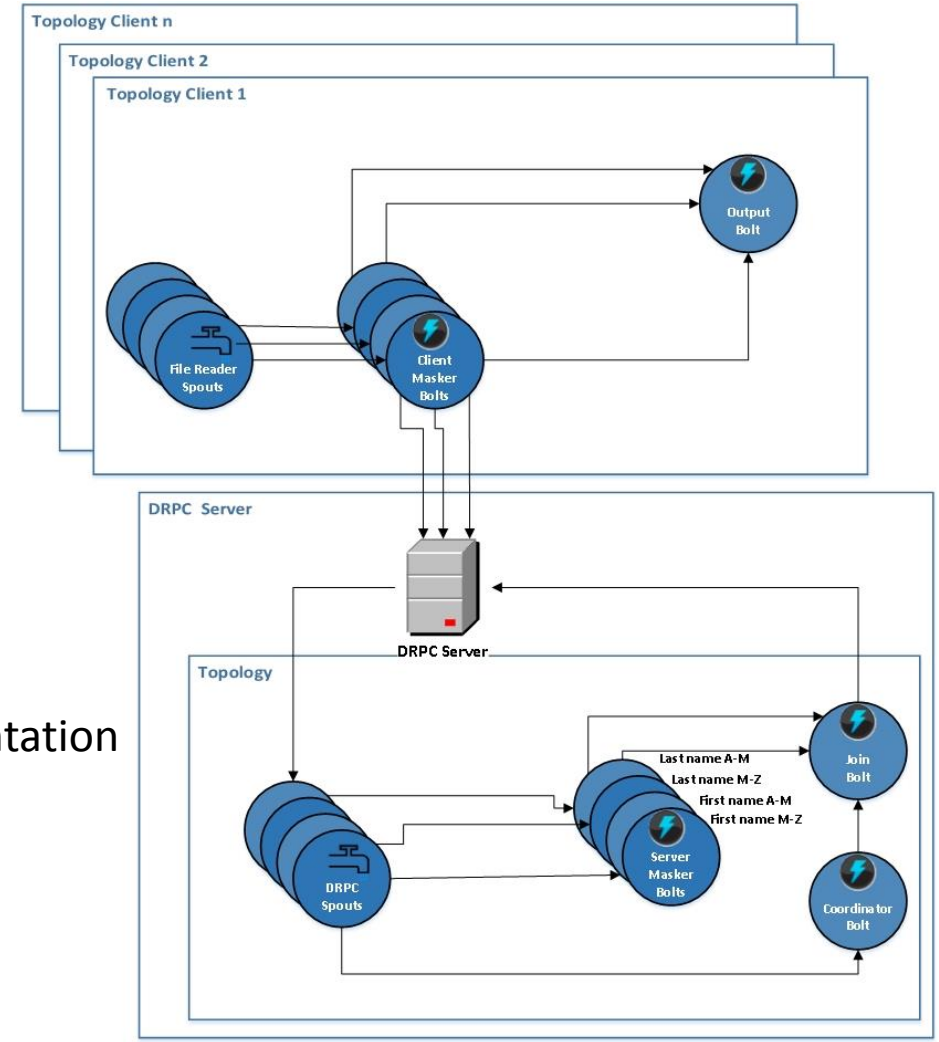
In-Situ Anonymization system



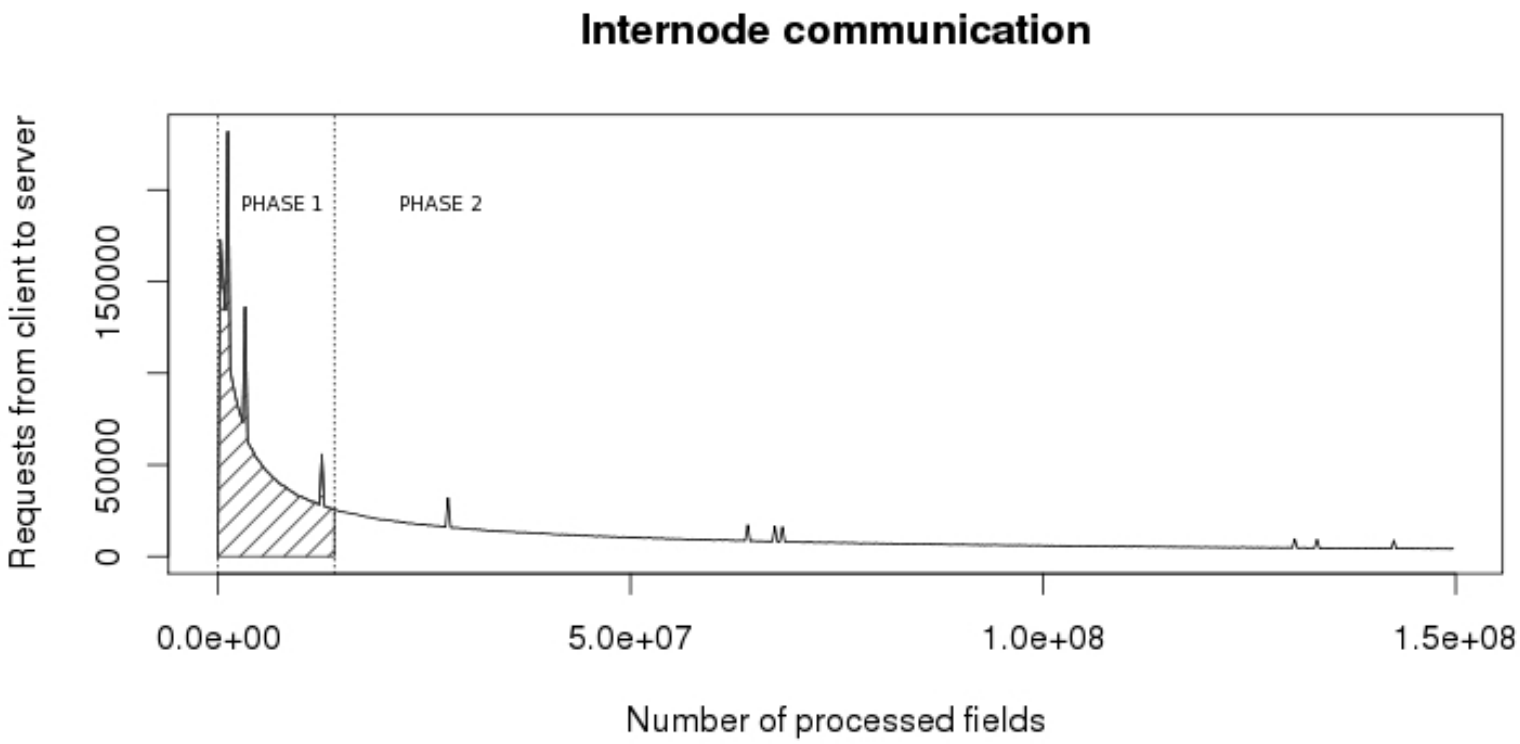
In-Situ Anonymization system over distributed streaming platform (e.g. STORM)



Logical vs Physical implementation



Client-Server communication over time



* Real-World measurement couldn't be shown due request from Vendor who are providing Data Masking solutions pending their laboratory testing

Conclusion

- Anonymization is becoming more important with growing volume of generated data and privacy protection concerns.
- Newer anonymization techniques are required to resist the background knowledge and linking attacks that are possible by versatile datasets available online.
- Masking techniques need to be built into processing systems capable of handling high volume and high velocity of Big Data.

Conclusion

- Our work presents a system suitable for in-situ anonymization of data in places where it was generated. It is appropriate for high volume of data with masking techniques' appropriate mapping data structures.
- Moreover, latencies and unexpected pauses are minimized to adhere to the requirements of high velocity real time data.
- Also, internode communication in a distributed environment is minimized

Future work

- Development of system doesn’t stops at this stage. In the future we plan to introduce:
  - new connectivity's (ODBC/JDBC, Native, message Tuples, etc.)
  - GUI for centralized definition of Data Masking rules
  - Development of new communication components to fully extend and utilize distributed frameworks