CASD-TeraLab
Secure Access to Big Data

UNECE 2017- Skopje

20 – 22 september

Saïd KHADRAOUI

CASD

UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

The Secure Data Access Center (CASD)

- Equipment to access and process **confidential** data
- Activities :
  - First French host for **administrative data**
  - Provider of services for the public and the private sectors
- Several Aspects :
  - Strong security including a **biometric authentification**
  - No data leak
  - Terminals-like remote access technology
  - Advice, support, maintenance
- Nearly **1000 researchers** in more than 400 research projects

CASD

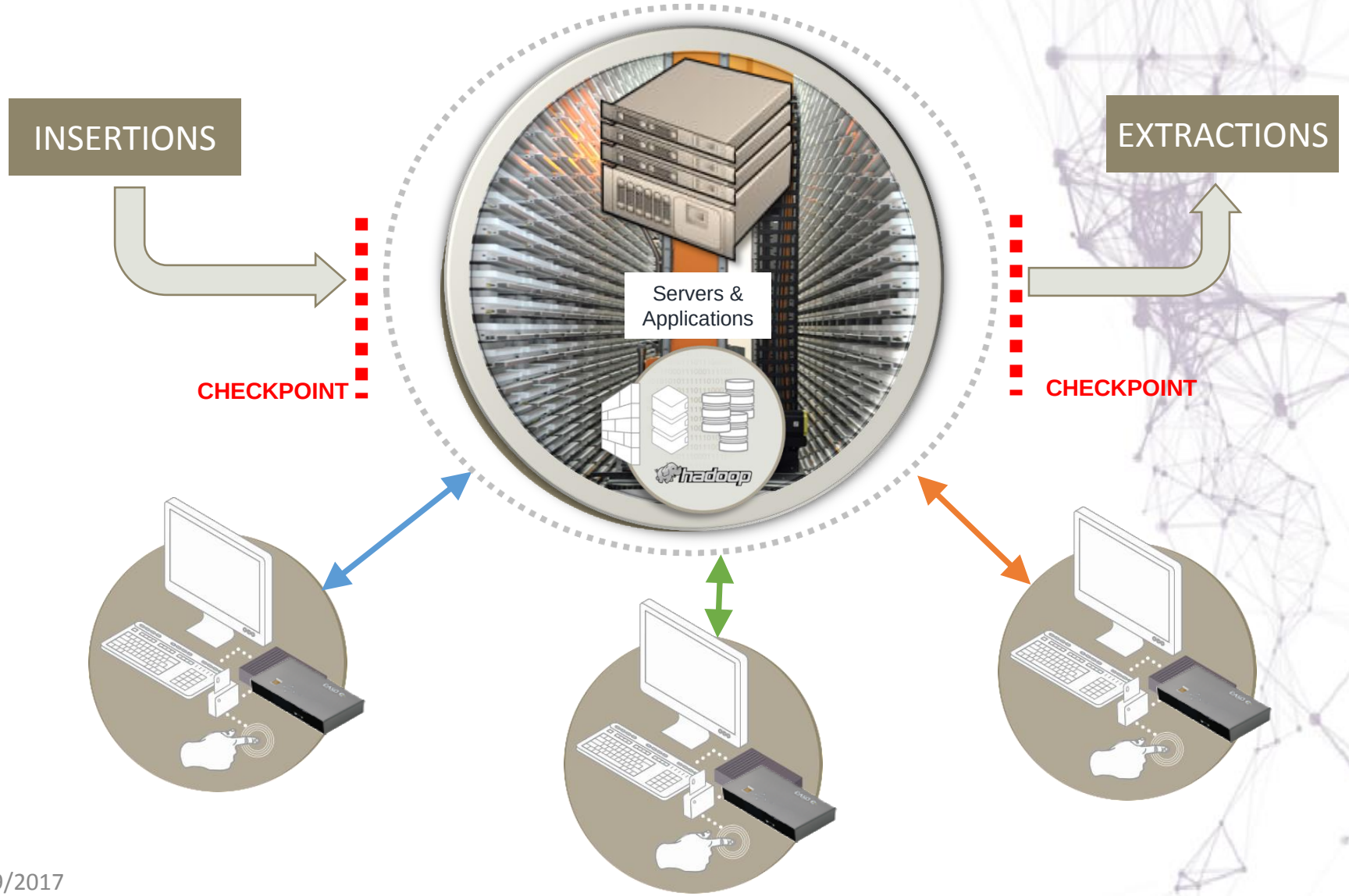
UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

Hermetically-sealed Chamber



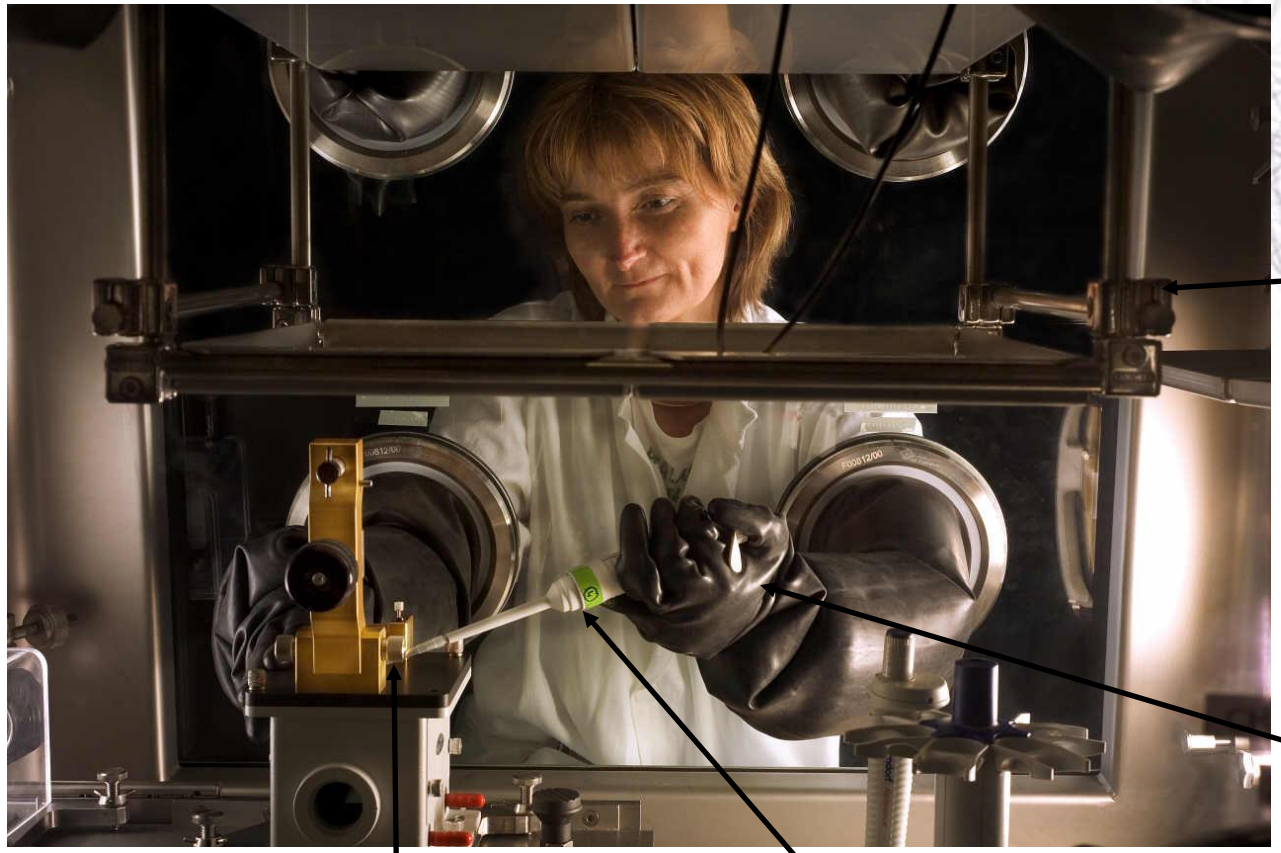
UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

The SD-BOX



UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

CASD-TeraLab

GENES Groupe des écoles nationales d'économie et de statistique

INSTITUT Mines-Télécom

- Publicly funded Big Data & Data Science platform
- Launched on 2014
- Partnership between :
  - GENES (*Groupe des Ecoles Nationales d'Economie et Statistiques*)
  - IMT (*Institut Mines Télécom*)
- Open to:
  - R&D and teaching projects, proof of concepts
  - Public and private sectors
- Everything for big data and data science:
  - Powerful and scalable infrastructure
  - Hadoop-based with all Hadoop tools, Spark
  - Extensive tools for scientists (Python, R, SAS, machine learning…)
- Turnkey solution with full support and maintenance

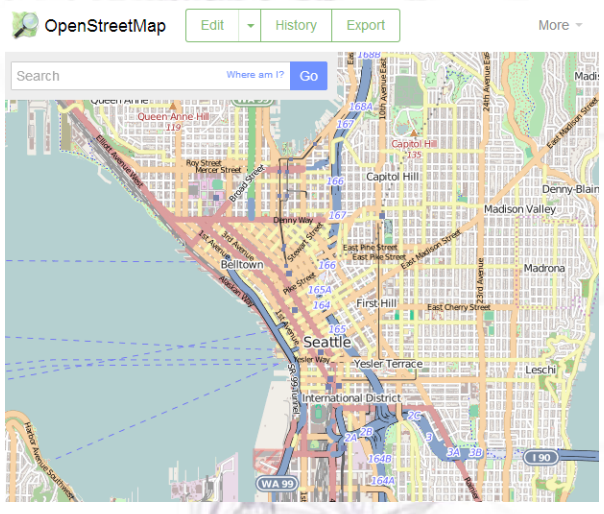
CASD

CASD-TERALAB,
A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

Tools available on the Clusters



- Scientific Tools :
  - Pig, Hive, Impala :
    - SQL database processing
  - R (RHadoop), Python :
    - Several existant libraries
    - Statistics, machine learning
  - Spark, Mllib :
    - Statistical libraries, machine learning
    - Regressions, clustering (K-means), classification
  - Open Street Map server

Said KHADRAOUI
Data-scientist
CASD

Data Disclosure

- Usually, individual data are obtained under a pledge of confidentiality; statistical entities aim at publishing as much information as possible without compromising the identities of their respondents .
- Trade-off between privacy protection and information loss.
- Cell suppression is a widely used technique
- At TeraLab and INSEE (French National Institute for Statistics and Economic Studies), tau-Argus is used.
- k-dimensional tables with k > 5 ??

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

Data:

- ADSD (Annual Declaration of Social Data) which is a file that contains several pieces of information that are transmitted to FISC administrations.
- INSEE, usually uses tau-Argus for its confidentiality needs, to anonymize this category of data.
- Plenty of explanatory variables that can be used by all sort of attackers to identify concerned individual.
- The response variable that we used to anonymize our data is *individual salary.*

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

❑ Tabulation:

- The very first step in the cell suppression process on tabular data is to construct the table to be processed from micro data.
- Constructing an array with the selected explanatory variables and including the hierarchies of these variables.
- Tau-Argus reaches its limits during the tabulation stage (it bugs when k >5).
- Within TeraLab, tests have been conducted to define the limits of tau-Argus software on a physical machine with sufficient GB of available RAM.

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

- The table below gives an idea of the performance of tau-Argus on a file of 1,306,620 rows (37.3 MB), on a dedicated physical machine:

Number of explanatory variabless	Secret secondaire (seconds) (« modular »)
4	3h27
>5	Can't work (bugs)

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

Different use cases

These "simple" examples help to better understand the approach described above.

Without hierarchy

SAS original data: p explanatory variables Vb,1, …. Vb,p and R the response variable.

The tabulation creates an array of p* explanatory variables V1, … Vp where p * ≤ p.

A single explanatory variable with m hierarchy levels

a. p =1, m=2:*

V has a hierarchical structure with two levels: V1 and V2 and V1 is the finest level and V2 is the most aggregated.

Example of micro-data:

V	V^1	V^2
2	0	0
2	0	0
10	2	0
9	1	1
9	1	1
9	1	1
15	3	1

UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

The searched table:

V	V^1	V^2	sum
2	0	0	2
10	2	0	1
9	1	1	3
15	3	1	1
NULL	0	0	2
NULL	2	0	1
NULL	1	1	3
NULL	3	1	1
NULL	NULL	0	3

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

b. p =1, m any integer:*

V, V^1,…, V^m where V^1 is the finest level and V^m is the most aggregated one.

The table we want to obtain has the following form where x denotes values and NULL denotes an empty field:

V	V^1	…	V^m	sum
x ⋮ x	x ⋮ x		x ⋮ x	x ⋮ x
NULL ⋮ NULL	x ⋮ x		x ⋮ x	x ⋮ x
NULL ⋮ NULL	NULL ⋮ NULL		x ⋮ x	x ⋮ x
NULL	NULL	NULL	NULL	x

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

c. Three explanatory variables, a single one with hierarchy

dads_idf_micro → dads_idf _micro_7 → dads_idf_micro_clean → dads_idf_micro_h → dads_idf_tab → dads_prim

Table	Rows number
dads_idf_micro	58 825 580
dads_idf_micro_clean	19 012 113
dads_idf_micro_7_h	19 459 448

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

❑ Primary cell suppression

To ensure primary secret, it is necessary to traverse all the cells of the table and to hide the cells which do not satisfy at least one of the following two criteria:

- Frequency rule: a cell must be constructed from at least k units (at INSEE ***k = 3***)
- Rule of dominance: the cell first contributor must not represent more than ***x%*** of the total value of the cell (to INSEE ***x = 85%***)

Explanation:

Let a cell $T_c = \sum_{i=1}^{n} w_i\, x_i$ where $x_1 \geq \cdots . \geq x_n$. x_i denotes the response value and w_i its associated weight. The cell is considered sensitive if $w_i \leq 3$ (frequency rule) or $x_1 > 0.85 \times T_c$ (dominance rule).

To ensure table protection with the specified levels of protection, a phase of additional cells suppression (secondary suppression) due to margins publication !!!

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

❑ Secondary Cell Suppression

A table is a data vector $\boldsymbol{a} = [a_1, \ldots, a_n]$ whose entries satisfy a given set of linear constraints known to a possible attacker,

My=b

$$lb_i \leq y_i \leq ub_i \forall i = 1 \ldots\ldots\ldots n \quad (1)$$

(1) models the whole a priori information on the table known to an attacker

let ***PS*** = $\{i_1, \ldots, i_p\}$ be the set of sensitive cells to be protected, as identified by the statistical office according to some criteria. For each sensitive cell $i_k (k = 1, \ldots, p)$, the statistical office provides three nonnegative values: $\boldsymbol{LPL_k}$, $\boldsymbol{UPL_k}$, and $\boldsymbol{SPL_k}$, the lower protection level, upper protection level, and sliding protection level, In typical applications, these values are computed as a certain percentage of the nominal value a_{i_k}.

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

A suppression pattern is a subset of cells, ***SUP*** $\subseteq$ ***{ 1, … , n},***corresponding to the unpublished cells. A consistent table with respect to a given suppression pattern SUP and to a given nominal table is a vector***y = [y*** $_1$***, . . . , y*** $_n$***]***satisfying:

$$\mathbf{My=b}$$

$$\mathbf{lb_i \leq y_i \leq ub_i} \quad \forall i \in SUP(2)$$

$$y_i = a_i \forall i \notin SUP$$

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

The Integer Linear programming model

We define :

$$LB_i := a_i - lb_i \geq 0$$
$$UB_i := ub_i - a_i \geq 0$$

that is, the range of feasible values for cell i known to the attacker is $[a_i - LB_i; a_i + UB_i]$

To obtain a mixed integer linear programming (MILP) model for the CSP, we introduce a binary variable $\boldsymbol{x_i = 1}$ for each cell $\boldsymbol{i}$, where $\boldsymbol{x_i = 1}$ if $\boldsymbol{i} \in$ SUP (suppressed), and $\boldsymbol{x_i = 0}$ otherwise (published). Clearly, we can fix $\boldsymbol{x_i = 0}$ for all cells that must be published (if any) and $a_i = 1$ for all cells that must be suppressed (sensitive cells). Using this set of variables, the model is of the form

$$min \sum_1^n w_i x_i \qquad (3)$$

s.t $x \in \{0,1\}^n$ and, for each sensitive cell $i_k (\boldsymbol{k = 1, \dots, p})$

UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

The mathematical model

Anew model inspired by Benders' decomposition (see, e.g., Nemhauser and Wolsey 1988); The idea is to use standard LP duality theory to project the auxiliary variables f^k and g^k (***k* = *1,…, p***) away from the model.
The new model is based on a characterization (reported in the Appendix on paper) of the vectors x that minimizes (3) so that the new model can be summarized to:

$min \sum_1^n w_i x_i (3)$

subject to $x \in \{0,1\}^n$and, for each sensitive cell $i_k(\boldsymbol{k} = \boldsymbol{1}, \dots, \boldsymbol{p})$

$$\sum_1^n (\alpha_i UB_i + \beta_i LB_i) x_i \geq UPL_k$$

$\forall\ (\alpha, \beta, \gamma) satisfiying\ (A.1)$ (5)

$$\sum_1^n (\alpha_i UB_i + \beta_i LB_i) x_i \geq LPL_k$$

$\forall\ (\alpha, \beta, \gamma) satisfiying\ (A.2)$ (6)

and

n

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

where (A.1) and (A.2) are linear systems governing the projection, as defined in the Appendix. Although these systems admit in general an infinite number of feasible solutions (α, β, γ) satisfying (A.1) and $(\alpha', \beta', \gamma')$ satisfying (A.2), it is well known that only a finite number of "extreme" solutions lead to undominated constraints (5)-(7); that is, these constraints are finitely many and hence define a convex polyhedron

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

Solving the model:

The model described in the previous section was resolved by implementing a meta-heuristic optimization algorithm which called tabu-search, the algorithm aims at finding the optimal solution of the LP problem. Tabu-Search is a meta-heuristic that guides a local heuristic search procedure to explore the solution space beyond local optimality .

- ✓ Adaptive memory
- ✓ This procedure has one parameter: the number of iterations, ***N***, which is used as the stopping criterion. The other arguments are a seed for initializing the random points chosen for cell suppression, and the name of the data file of dads.

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

The part of the MILP solution that is determined by tabu-search is the subset of integer variables ***x*** in Equation 3. The data structure representing a solution is therefore an ***n***-dimensional vector of integers, $\boldsymbol{X} = (x_1, \dots, x_n)$, ***n*** is the number of cells to be suppressed, it is therefore variable and can change through the progress of the tabu search.

Two search possibilities:

- ✓ *Breadth first:* all the neighbor solutions are checked, and the best is returned.
- ✓ *Depth first:* as soon as a non-tabu solution better than the current solution is found, it is returned.

Results obtained for these two strategies are rather similar, but there seems to be a slight advantage to breadth first, which is the Strategy that adopted is this paper.

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

Cluster archeticture within TeraLab/CASD:

The Technical specifications of the cluster are listed below:

- ✓ 108 GB of RAM
- ✓ 18 cores for computational processors (Dell PowerEdge R630).
- ✓ 8 TB of disk storage.
- ✓ 10 Giga Bytes for Ethernet network.

UNECE Speech
hh:mm

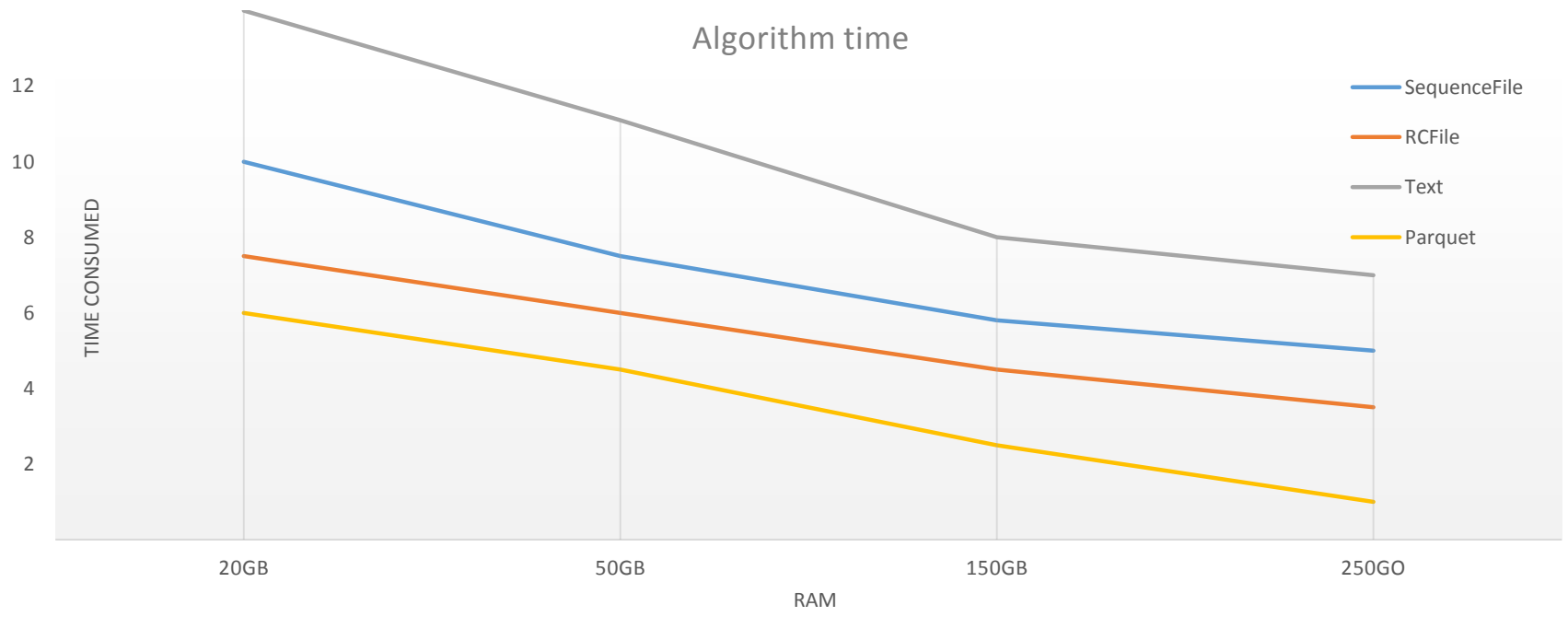
CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

Tabular Data Anonymization Process

Algorithm time amelioration:



Said KHADRAOUI
Data-scientist
CASD

European Level

- Involvement in European projects
- Open to share experience with similar European and international platforms

UNECE Speech
hh:mm

CASD-TERALAB, A Secure Remote Access System To Confidential Big Data.

FRANCE

Said KHADRAOUI
Data-scientist
CASD

THANK YOU

For more information:

- www.teralab-datascience.fr
- www.casd.eu
- Said.khadraoui@casd.eu
- datascience@casd.eu
- service@casd.eu