Sécuristation du Cloud

Schémas de recherche sur données chiffrées avancés

Alexandre Anzala-Yamajako

Laboratoire de Cryptologie
Thales Communications & Security

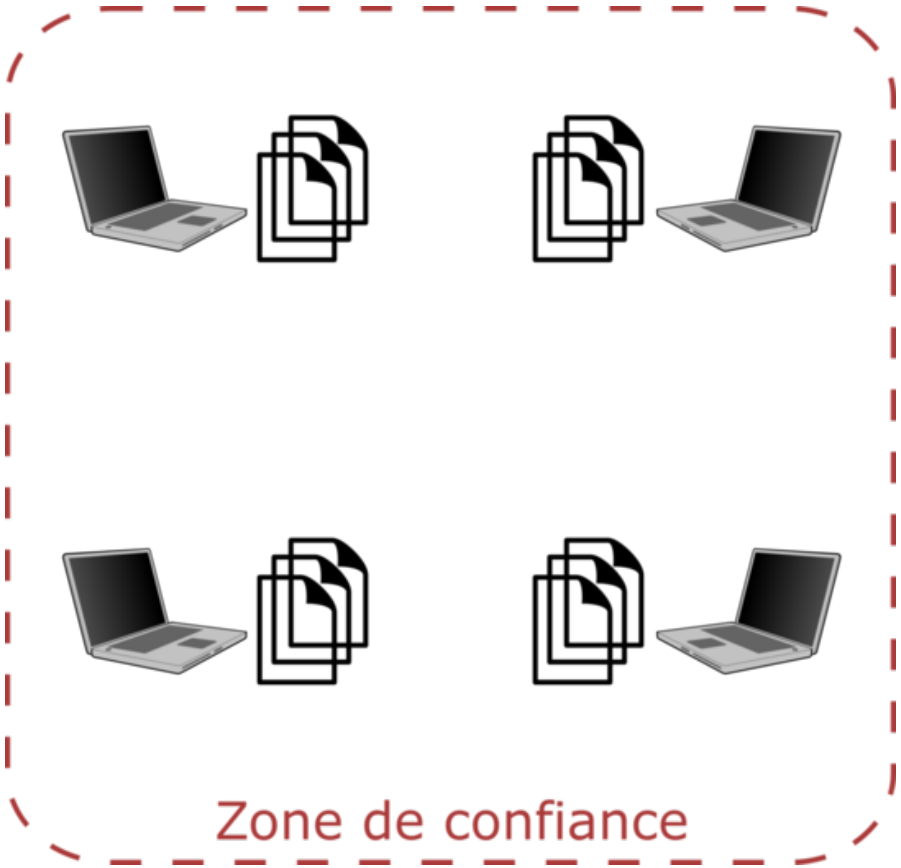
09 Avril 2015

Contexte

Aujourd'hui une entité est capable de maîtriser un périmètre de sécurité interne .

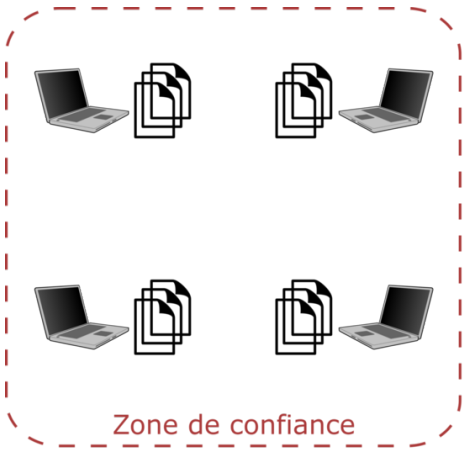
Contexte

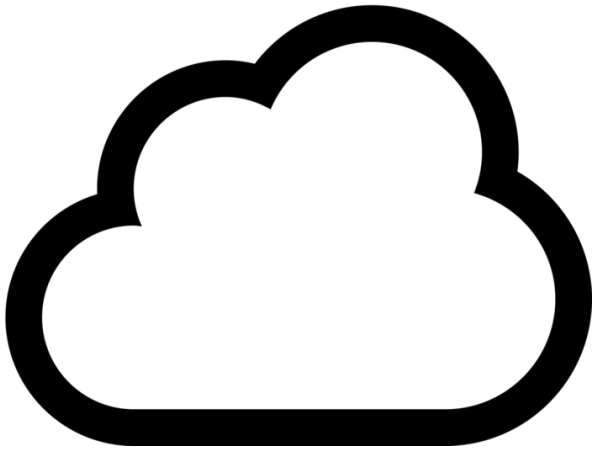
Cependant le volume des données numériques produites ne cesse de croître et beaucoup sont contraintes d'externaliser leur stockage de données .



Contexte

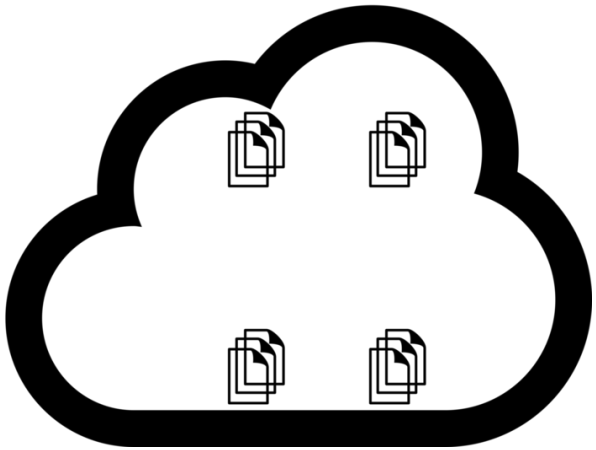
De nombreux opérateurs de Cloud sont prêts à accueillir ces données.



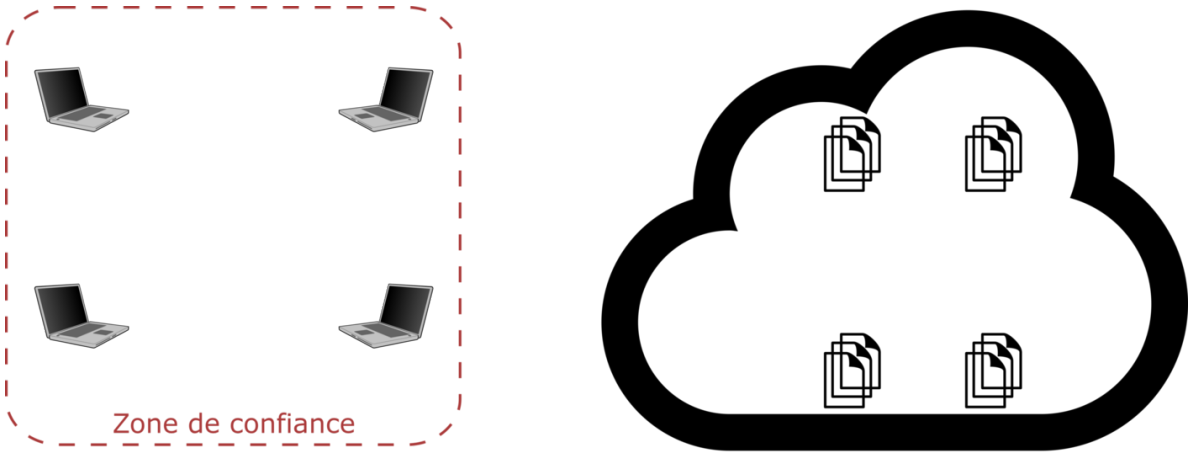


Contexte - Fonctionnalités sans Sécurité

Une partie des opérateurs de Cloud fonctionnent aujourd'hui sur un modèle "orienté Service ".



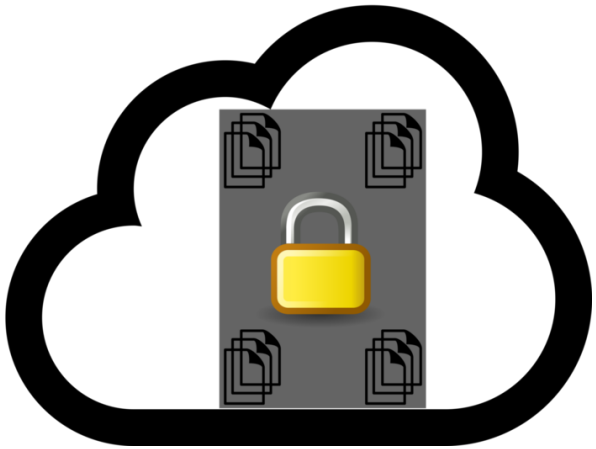
Contexte - Fonctionnalités sans Sécurité



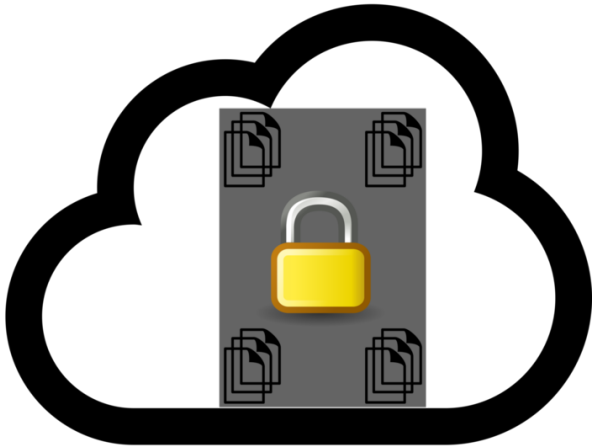
+ Fonctionnalité
- Sécurité

Contexte - Sécurité sans Fonctionnalités

D'autres choisissent d'offrir de fortes garanties de sécurité à leurs clients.



Contexte - Sécurité sans Fonctionnalités



+ Sécurité

- Fonctionnalité

Objectif

Développer une solution de recherche évoluées sur données chiffrées répondant à un compromis différent :

Sécurité

Garantir que l'opérateur de Cloud n'apprenne ni le contenu des requêtes ni le contenu des réponses via une preuve de sécurité.

Fonctionnalité

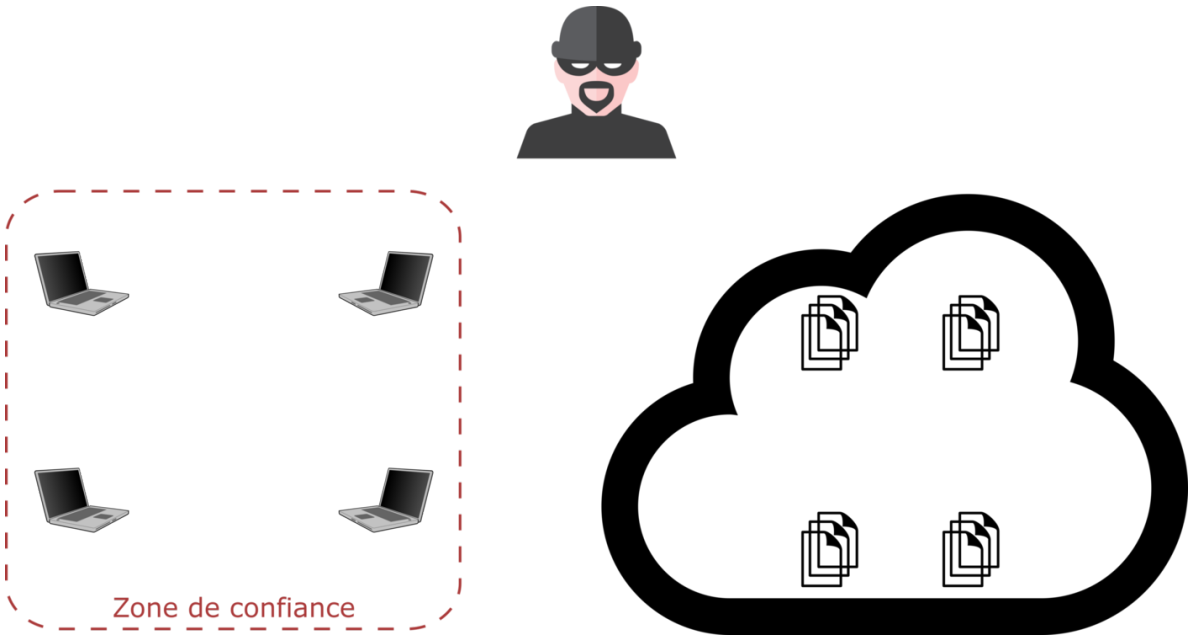
Offrir à l'utilisateur un service de recherche évolué par mots-clés .

Efficacité

Assurer un niveau de performances comparable à celui d'une solution non sécurisée .

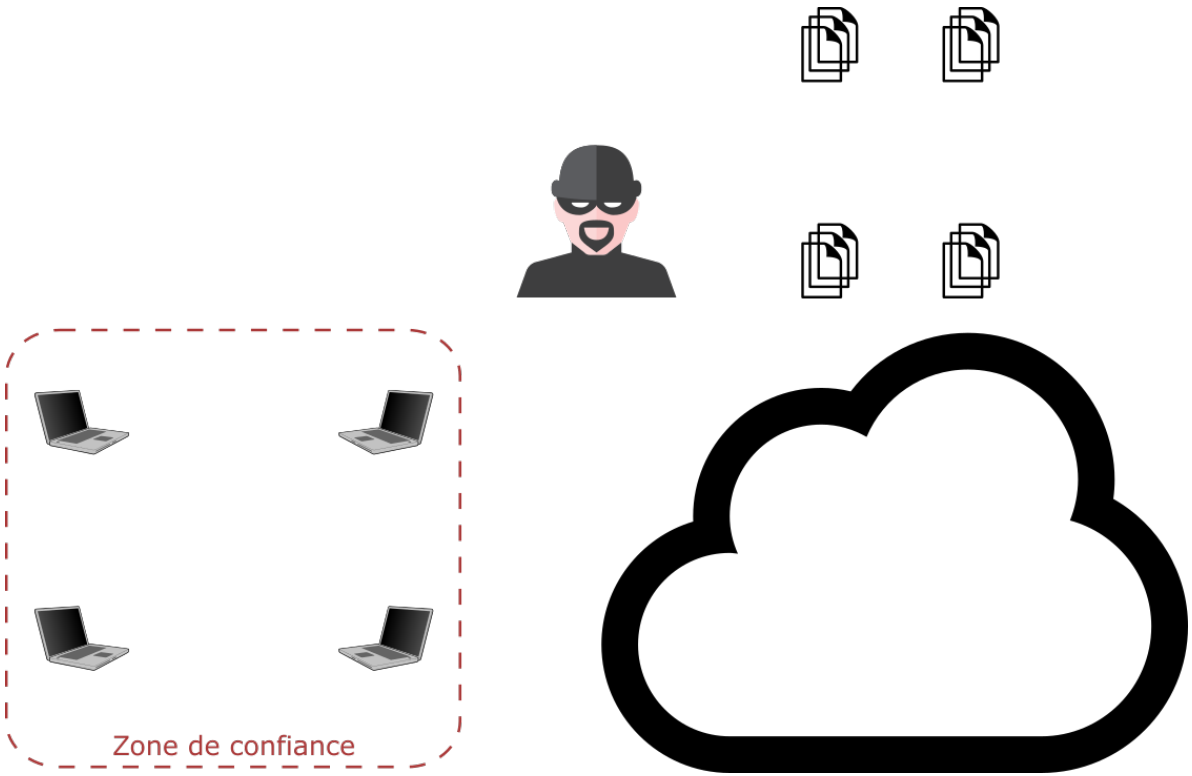
Menaces - Attaquants extérieurs 1

Même si l'on fait confiance à son opérateur de Cloud, ce dernier peut être victime d'une attaque informatique .



Menaces - Attaquants extérieurs 2

Nos données sensibles sont alors révélées .



Menaces - opérateur de Cloud public semi-honnête

Le modèle d'attaque pertinent est de considérer que l'on veut protéger nos données de l'opérateur de Cloud qui en assure le stockage.

Menaces - opérateur de Cloud privé semi-honnête

On peut utiliser aussi cette solution afin de réduire le niveau de sensibilité d'un opérateur de Cloud interne.

Acteurs

Entités manipulant de gros volumes de données

- Grandes entreprises ;
- Hôpitaux ;
- Administrations ;
- Organisations militaires.

Contrainte de sécurité dans le cas de données sensibles

- Stratégiques : affaires, brevets ;
- Confidentielles : données personnelles, classifiées.

Plan

Plan

Paramètres & Notations

Paramètres

- $|DB|$ nombre de documents dans la base de données ;
- $|W|$ nombre de mots-clés distincts ;
- $N = \sum_w |DB(w)|$ nombre de couples document/mot-clé.

Format des recherches évoluées

Une requête évoluée $\psi(\bar{w})$ est définie par un ensemble de mots-clés $\bar{w} \in W$ et une formule booléenne ψ sur $\bar{w}$.

On note $DB(\psi(\bar{w}))$ l'ensemble des documents qui correspondent à la requête $\psi(\bar{w})$ dans DB.

Algorithmes

Ces deux algorithmes sont des protocoles impliquant le client et le serveur.

Initialisation

`EDBSetup` prend en entrée une base de donnée `DB` et renvoie une clé K ainsi qu'une structure `EDB`

$$(K, EDB) \leftarrow EDBSetup(DB)$$

Recherche

`Search` prend en entrée la clé K, $\psi(\bar{w})$ et `EDB` et renvoie les indices $DB(\psi(\bar{w}))$

$$DB(\psi(\bar{w})) \leftarrow Search(K, \psi(\bar{w}), EDB)$$

Plan

Modèle de Sécurité - Introduction

Être sûr n'a de sens que dans un modèle de sécurité.

Fonction de fuite

On donne la liste précise des informations qui fuient au niveau du serveur via une fonction de fuite qui permet quantifier précisément cette fuite.

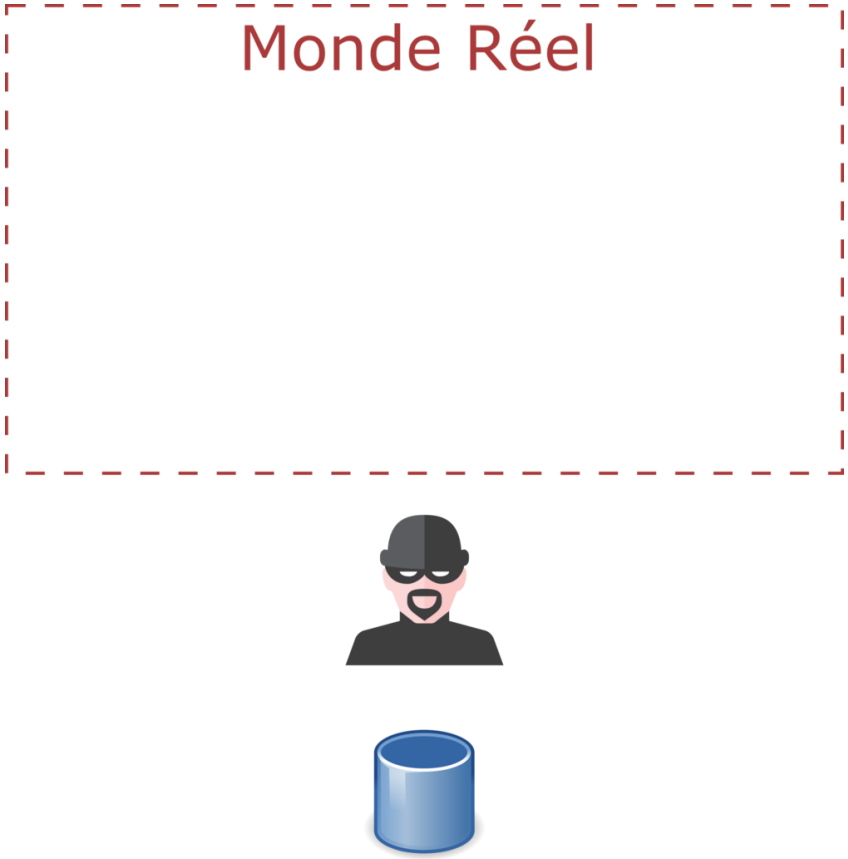
Principe du modèle

L'objectif est de prouver l'exhaustivité de la fonction de fuite . On compare

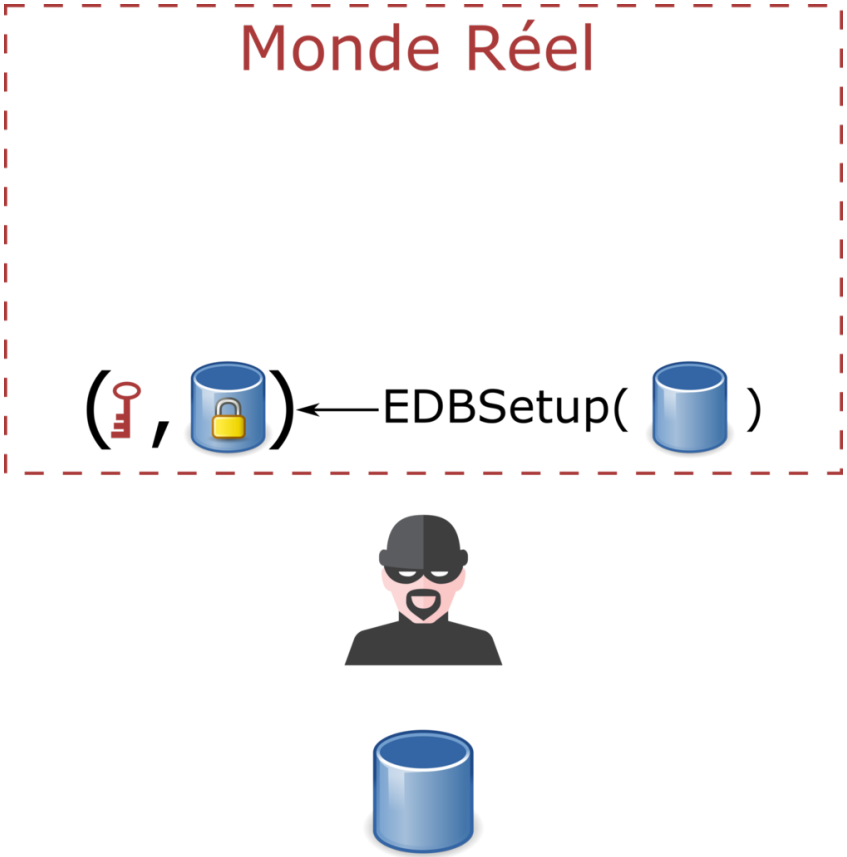
- un monde où l'exécution des protocoles est normale ;
- un monde où on simule la vue de l'attaquant à partir de la fonction de fuite.

Monde Réel

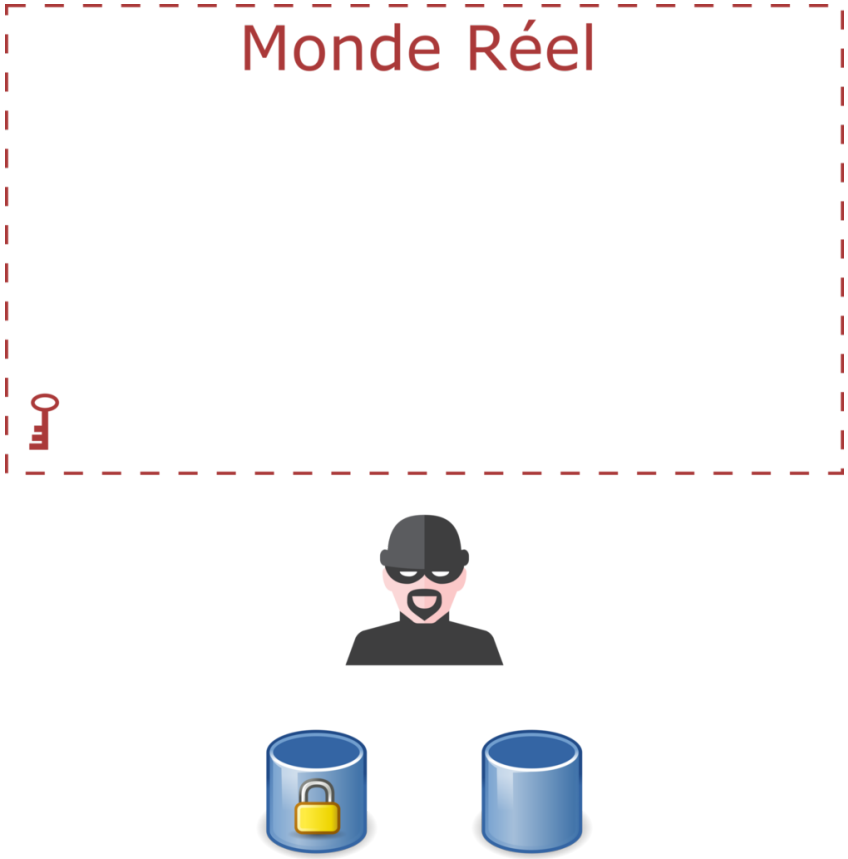
Monde Réel



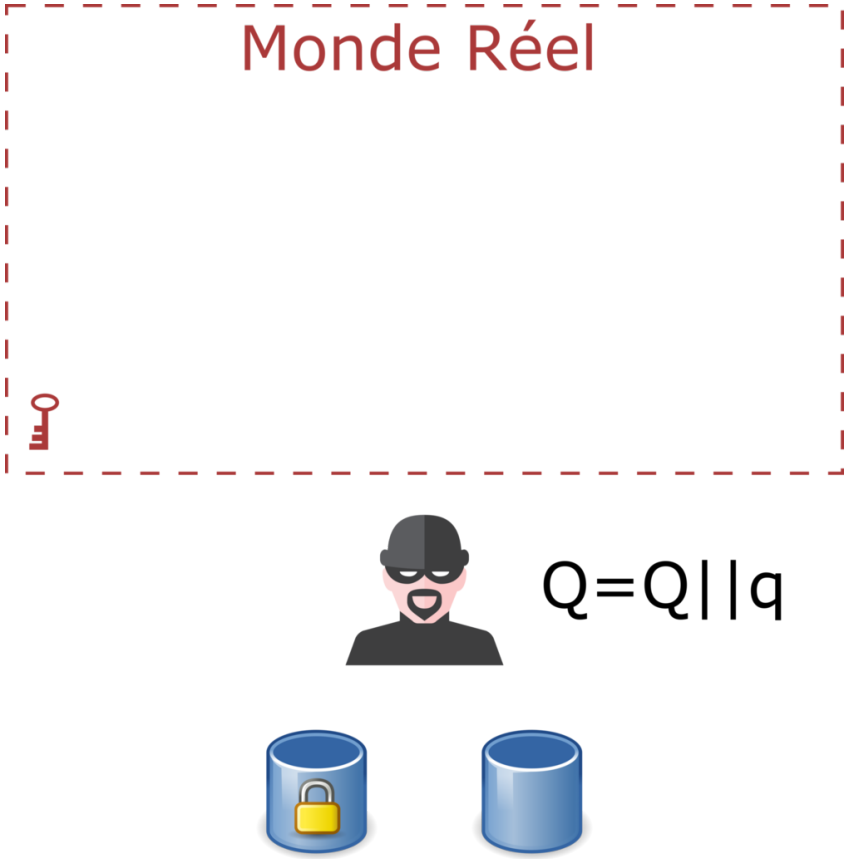
Monde Réel



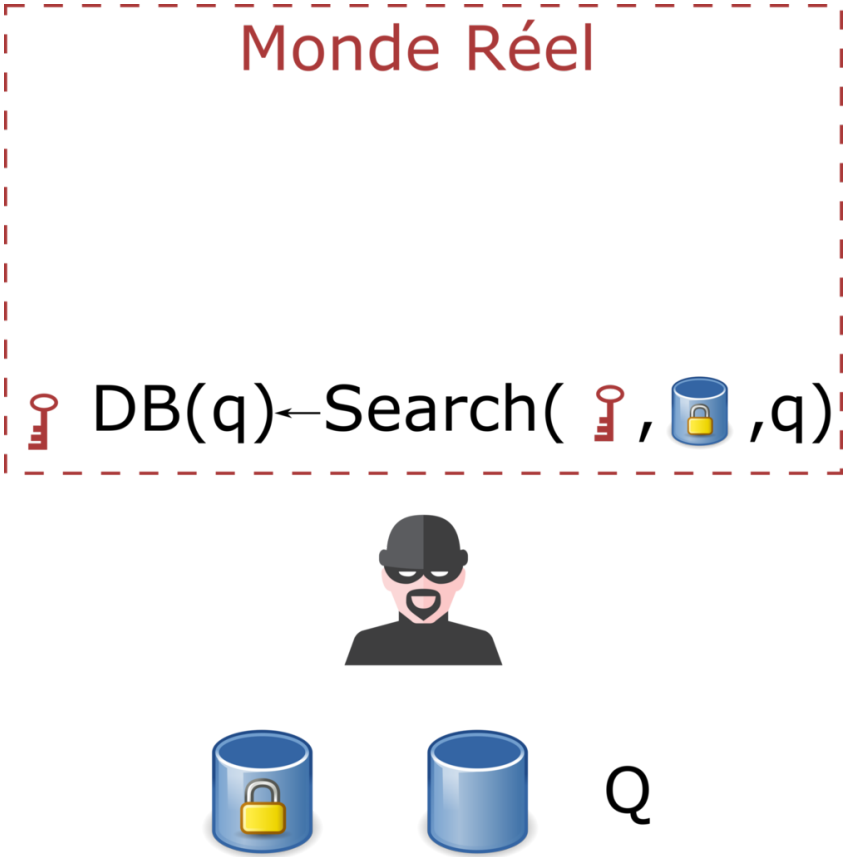
Monde Réel



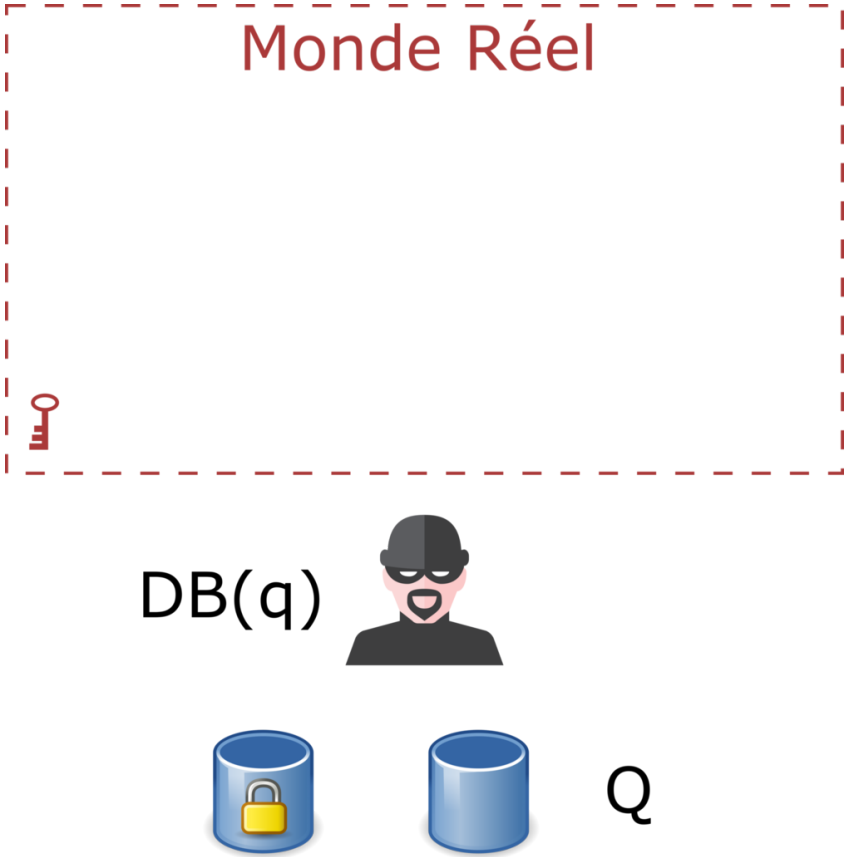
Monde Réel



Monde Réel

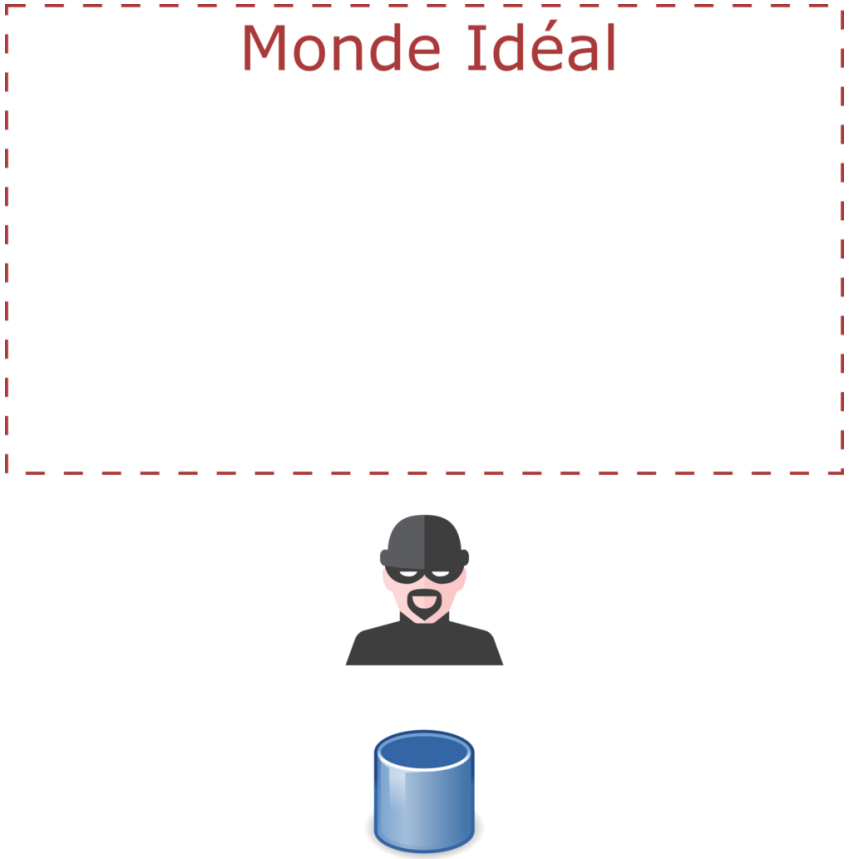


Monde Réel

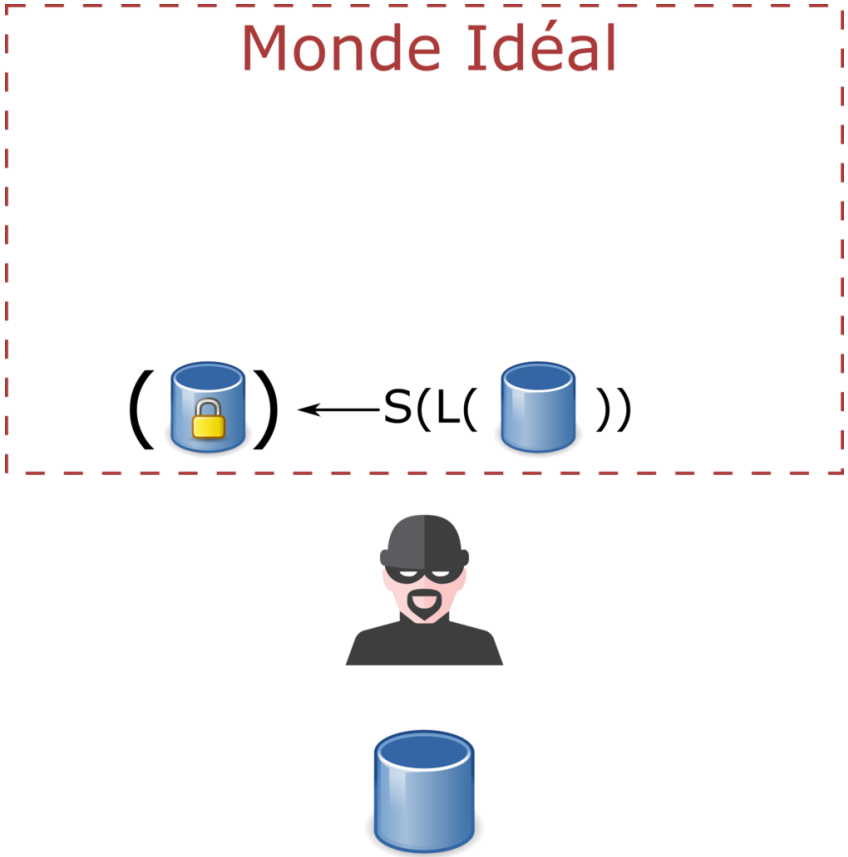


Monde Idéal

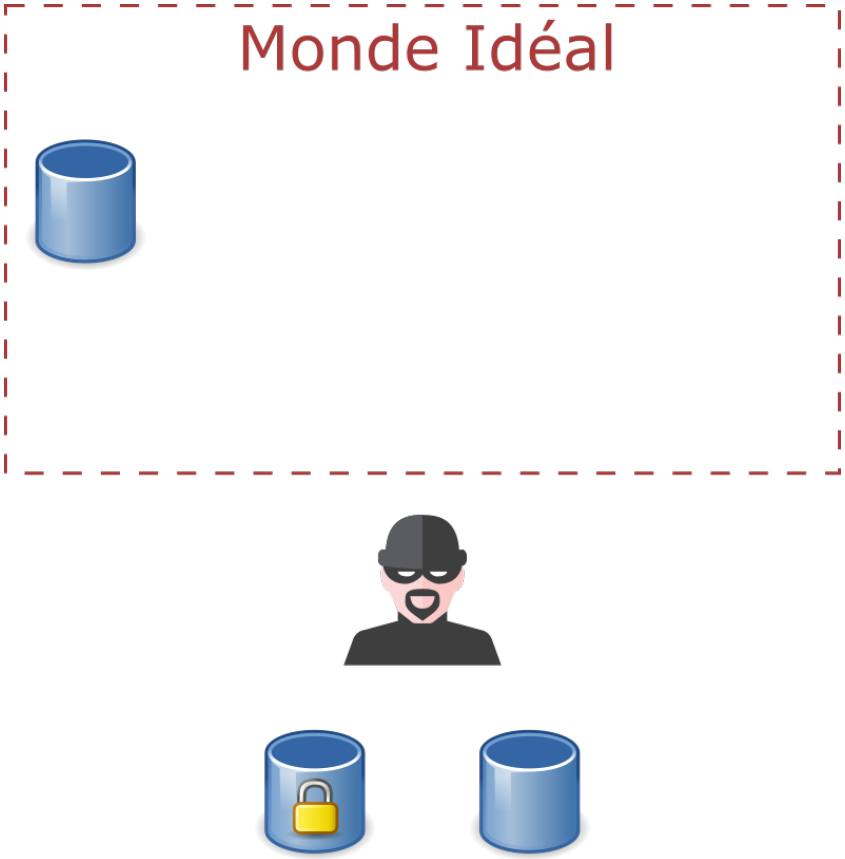
Monde Idéal



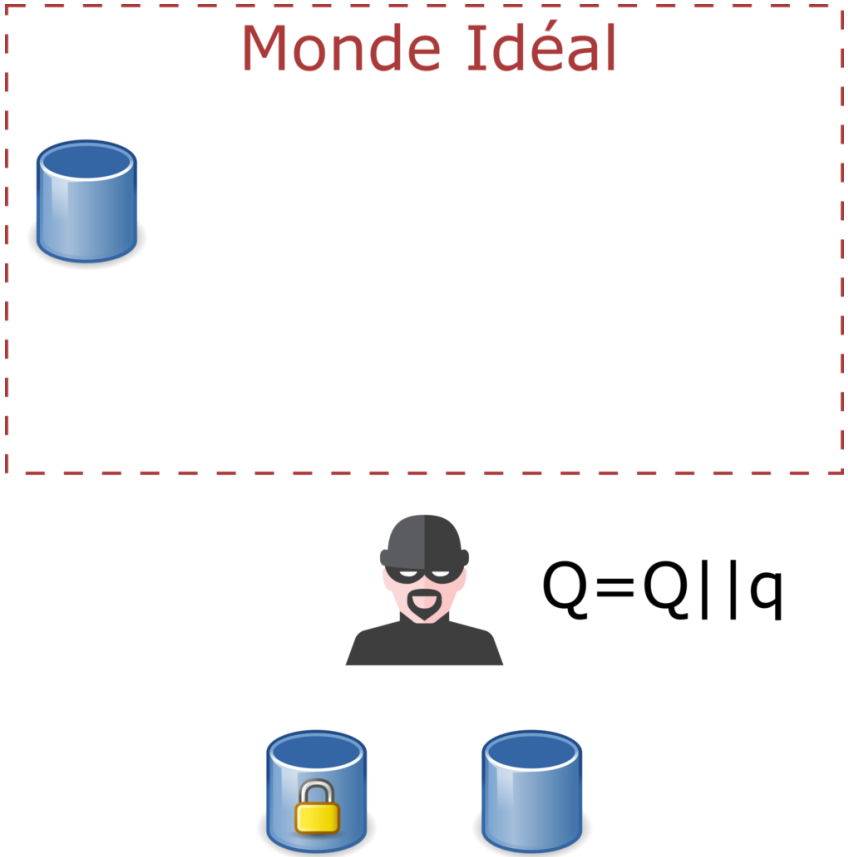
Monde Idéal



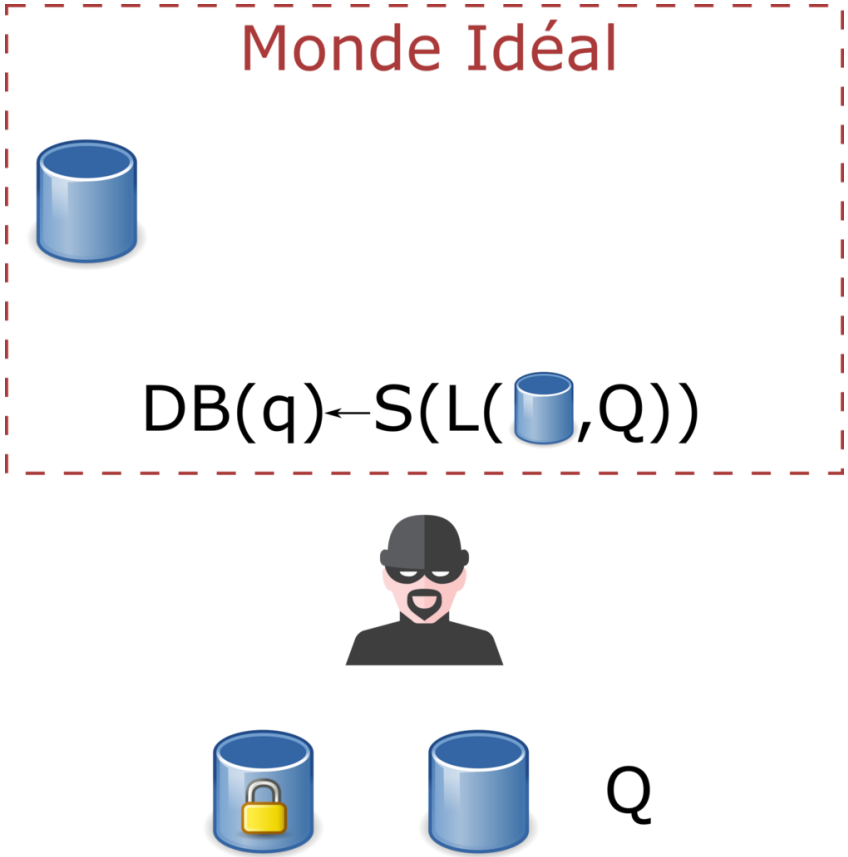
Monde Idéal



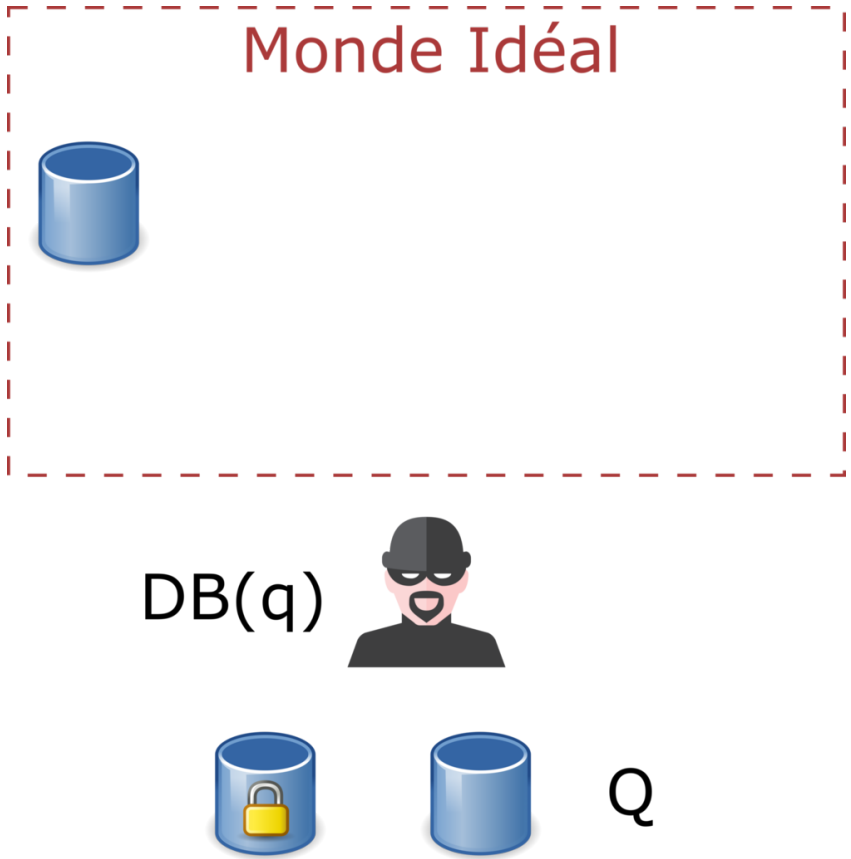
Monde Idéal



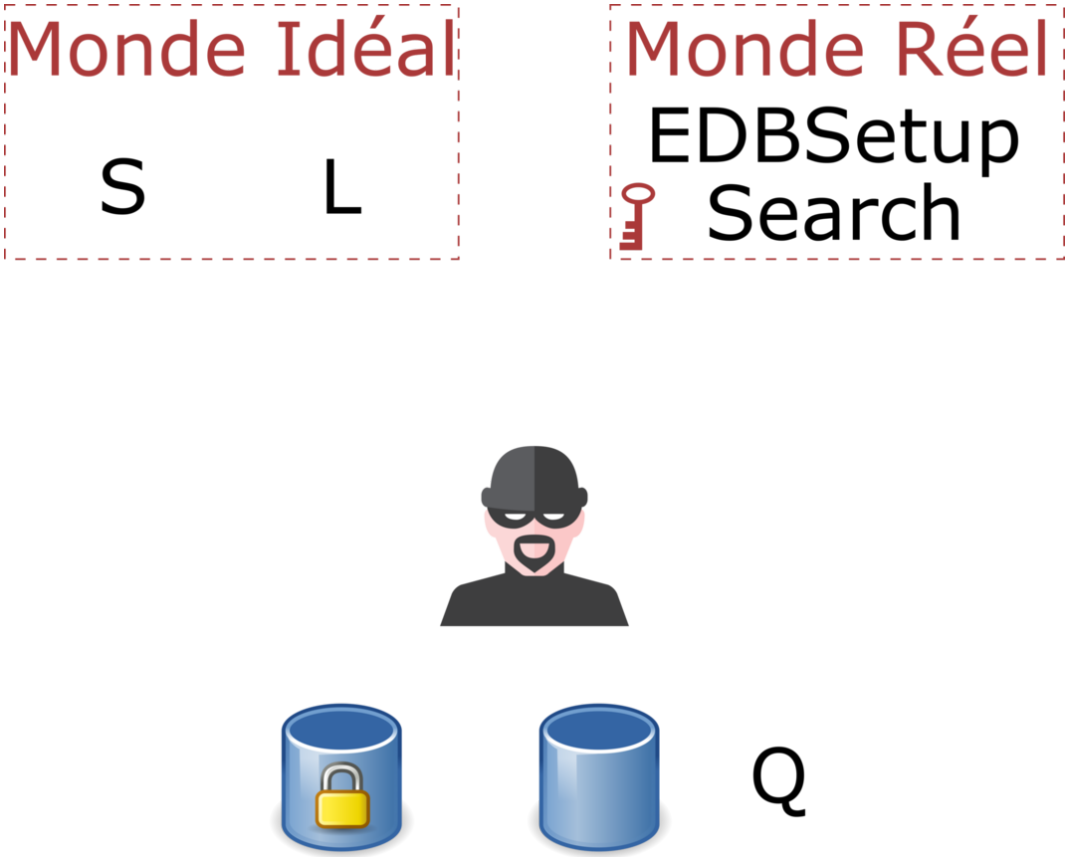
Monde Idéal



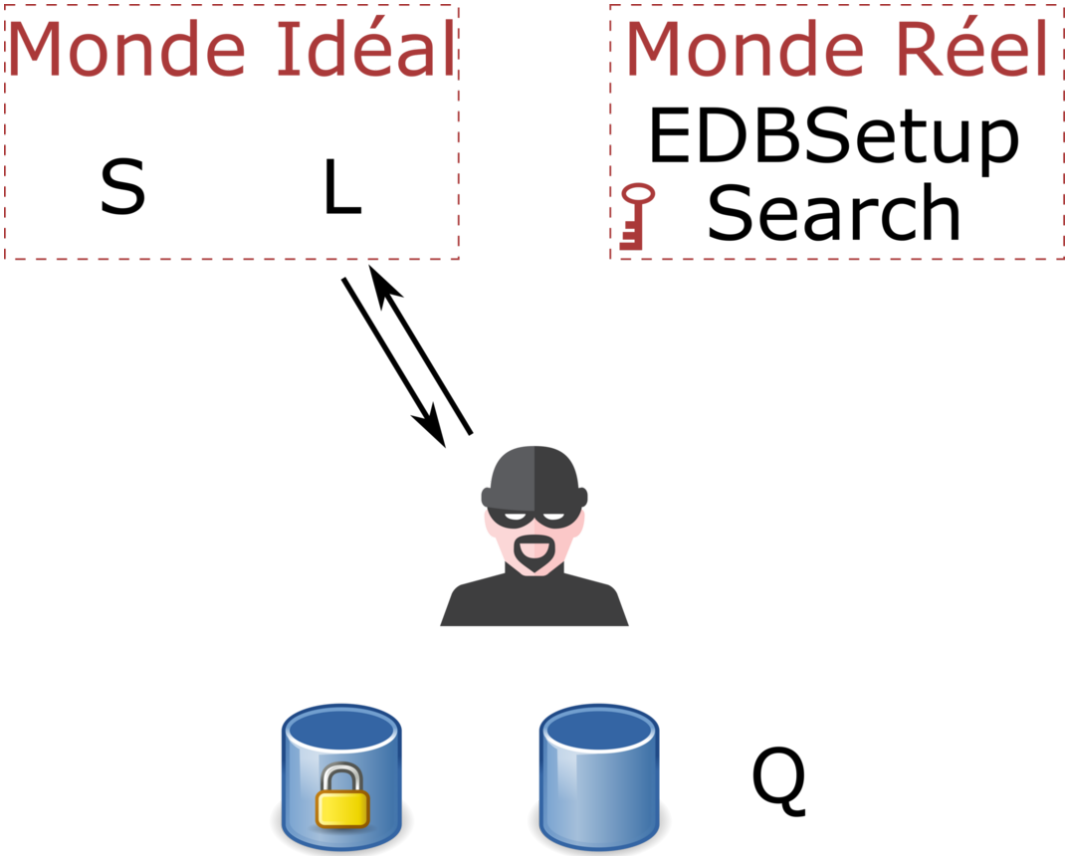
Monde Idéal



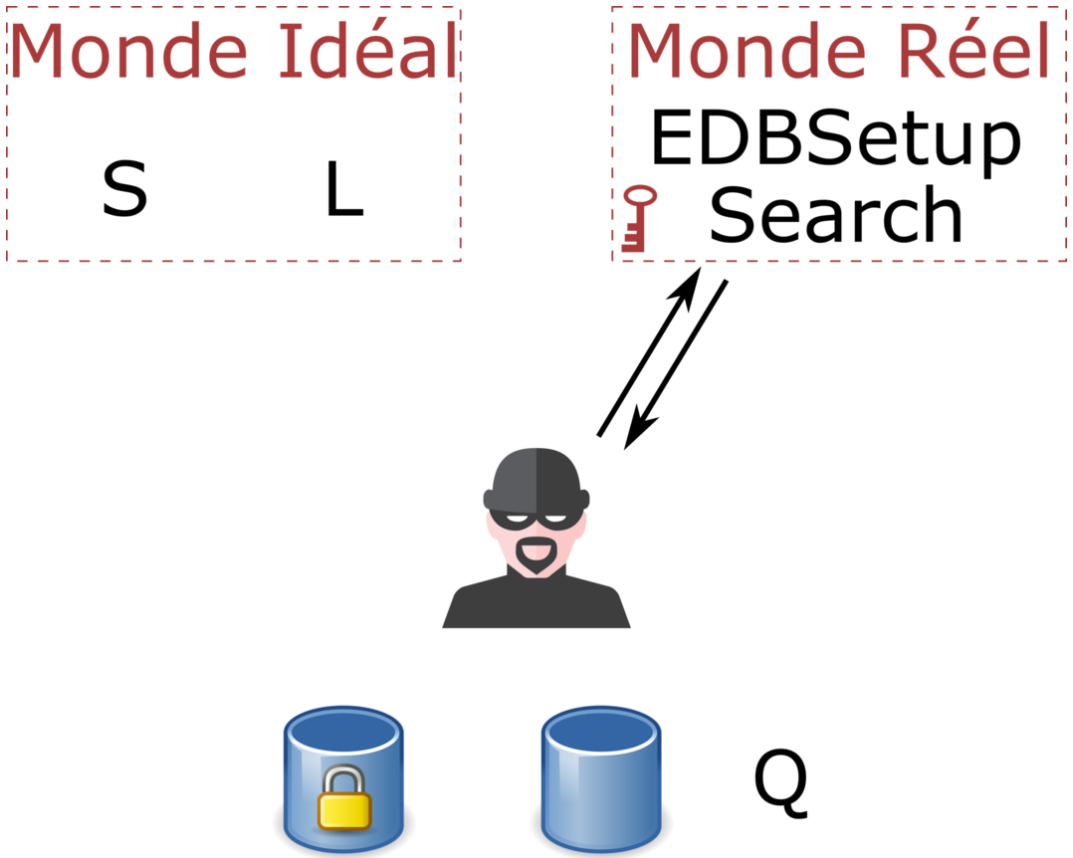
Objectif de l'attaquant



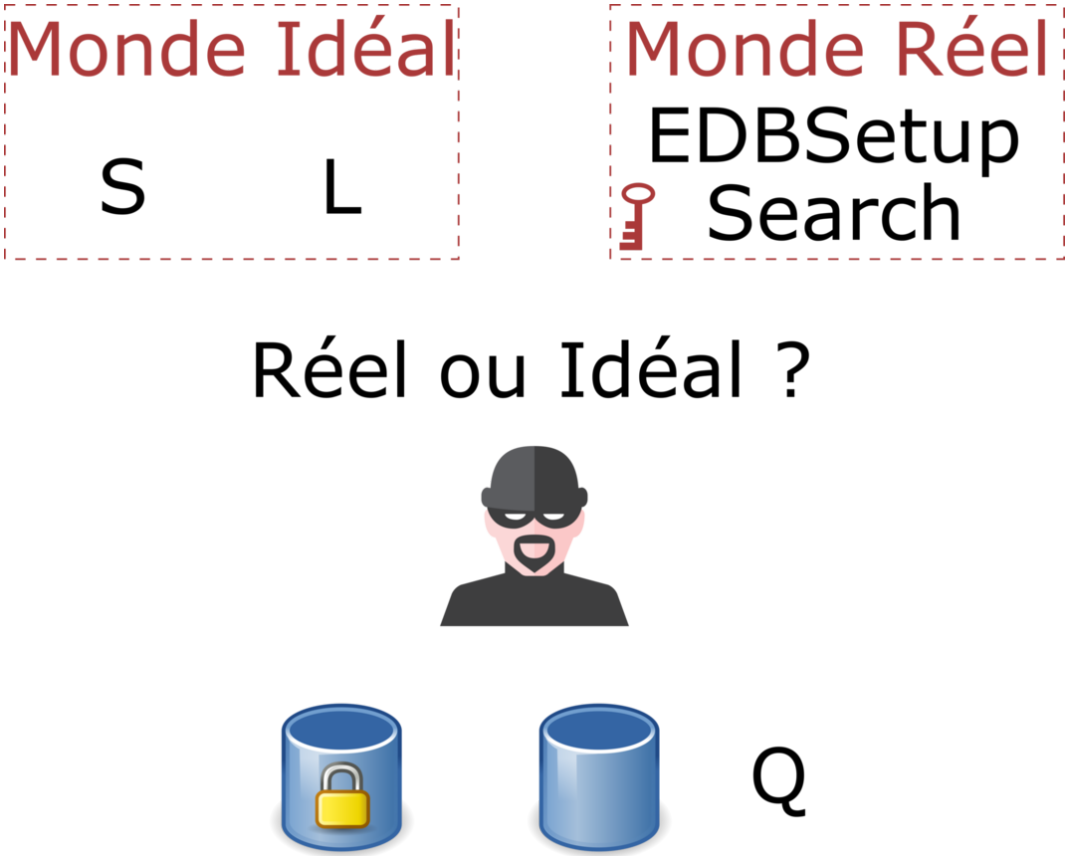
Objectif de l'attaquant



Objectif de l'attaquant



Objectif de l'attaquant

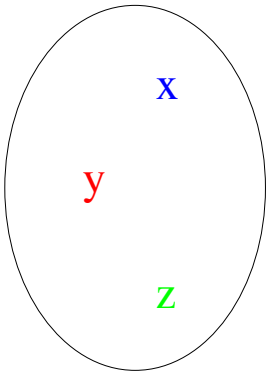


Plan

Principe du filtre de Bloom

Le filtre de Bloom d'après [Bloom, 1970], est une méthode de représentation compacte d'un ensemble permettant de vérifier l'appartenance en temps constant.

Ensemble



$n = \#\{x, y, z\} = 3$

Fonctions de hachage

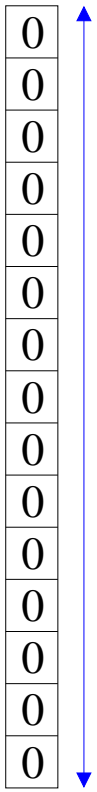
h_1

h_2

h_3

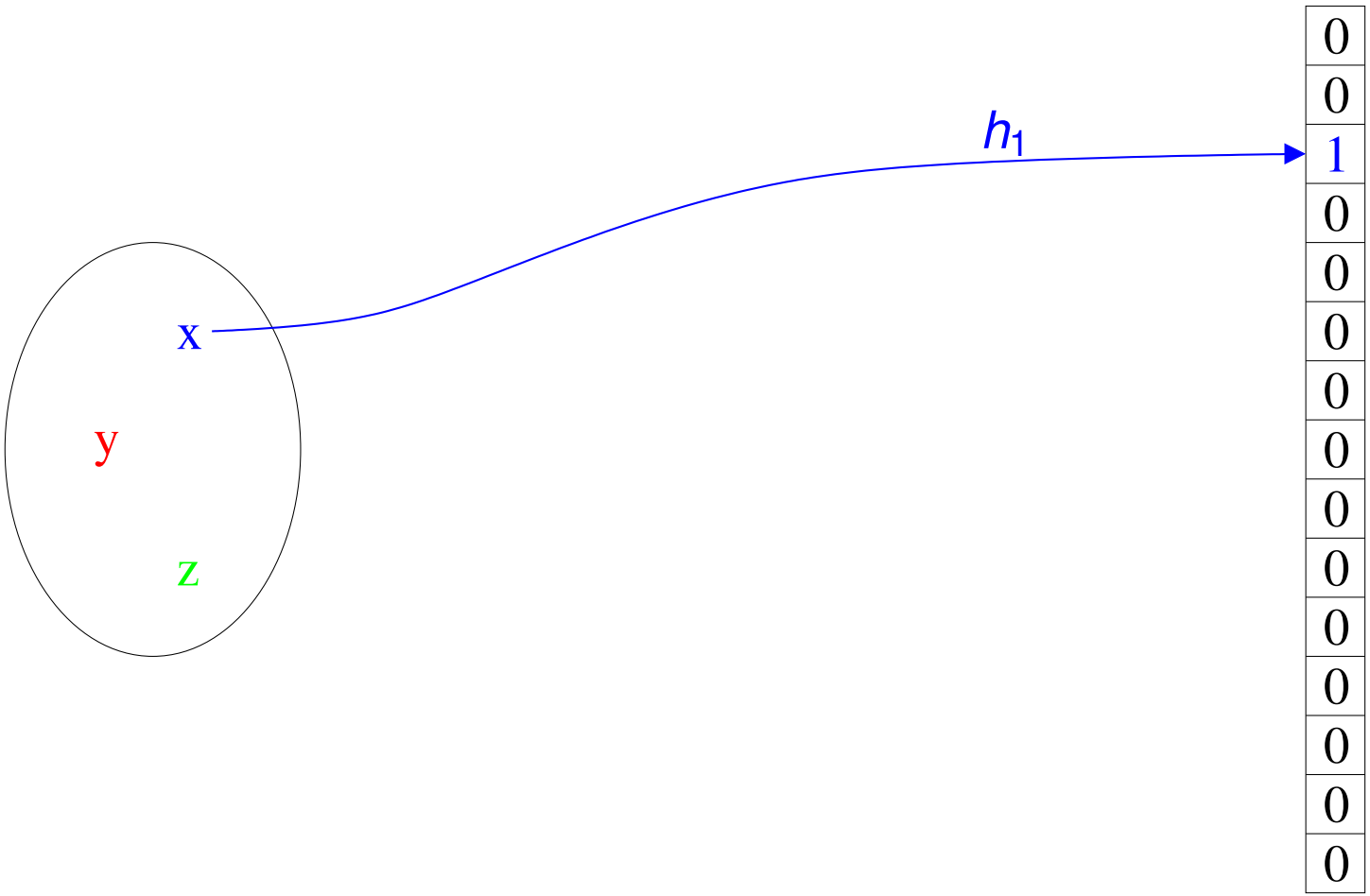
$k = \#\{h_1, h_2, h_3\} = 3$

Filtre de Bloom

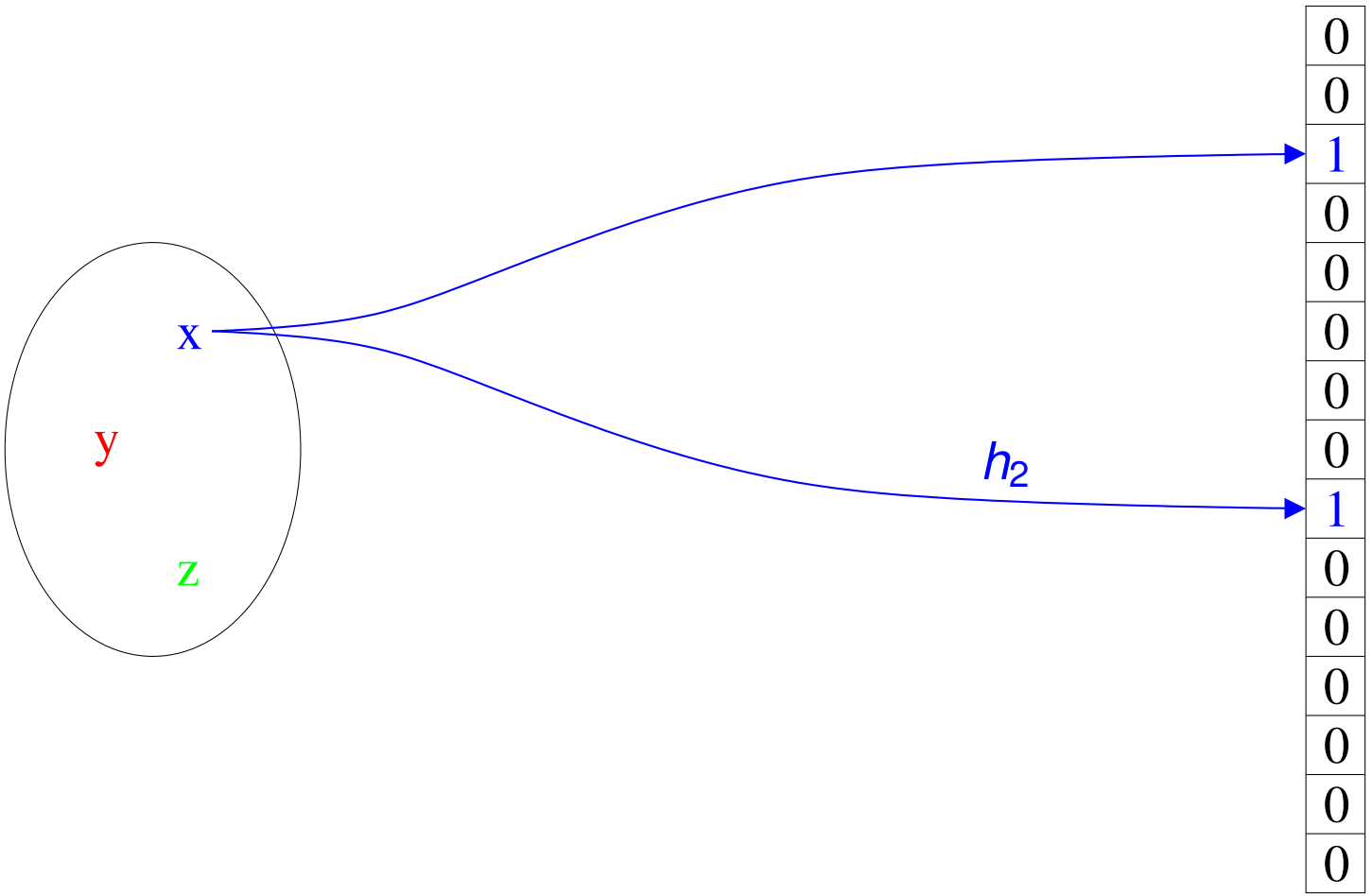


$m = 15$

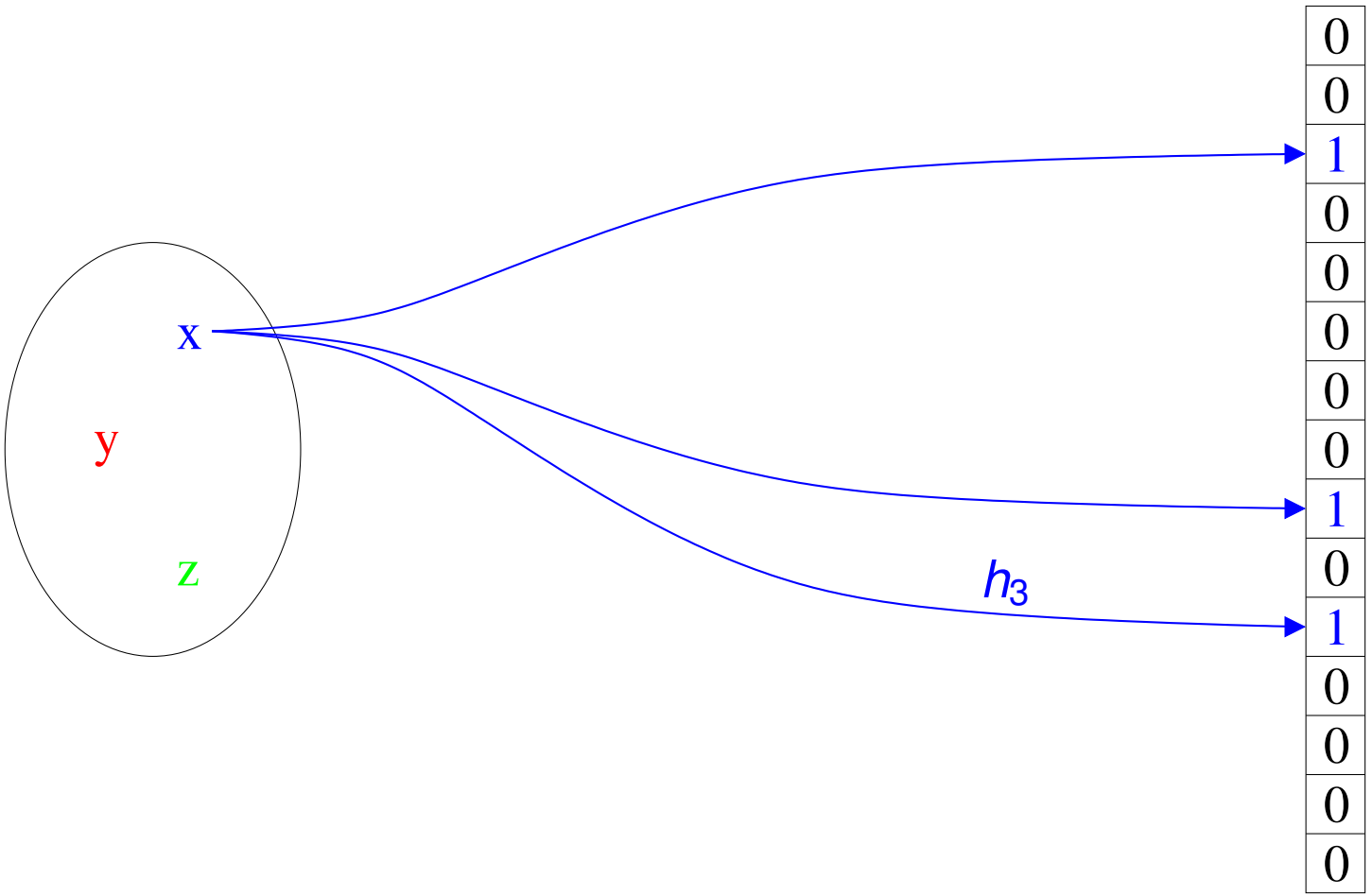
Insertion de l'élément x



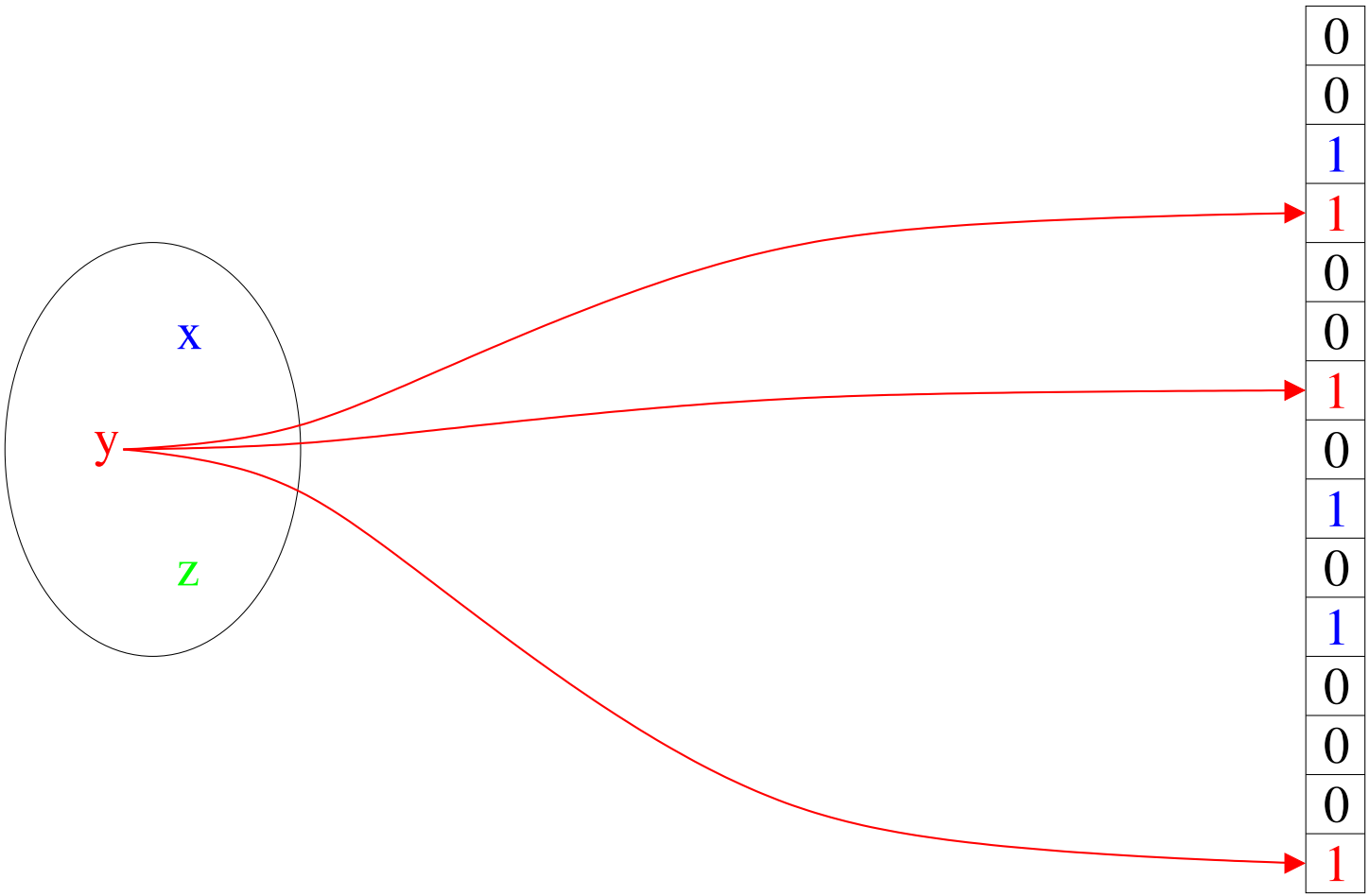
Insertion de l'élément x



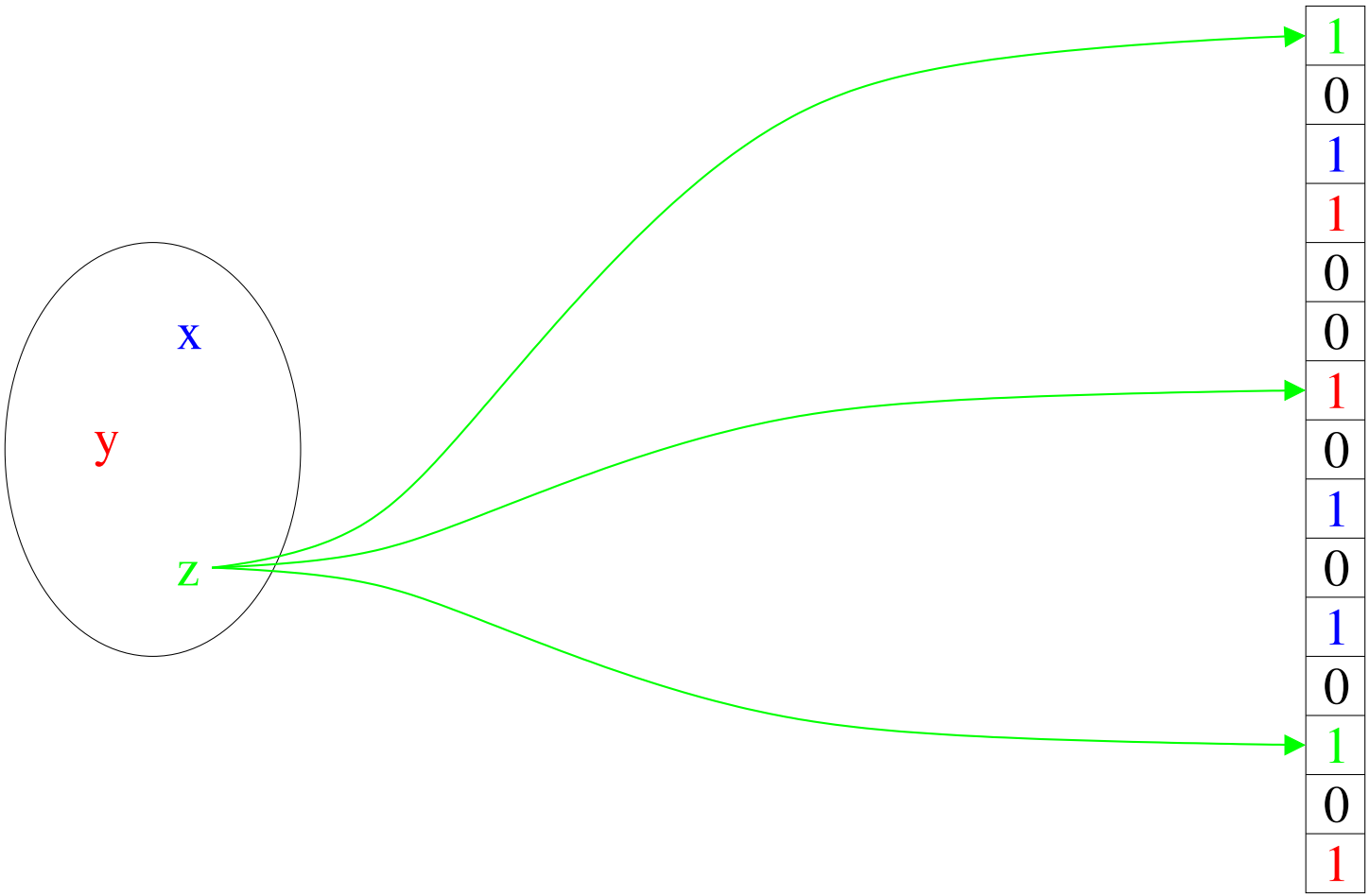
Insertion de l'élément x



Insertion de l'élément y

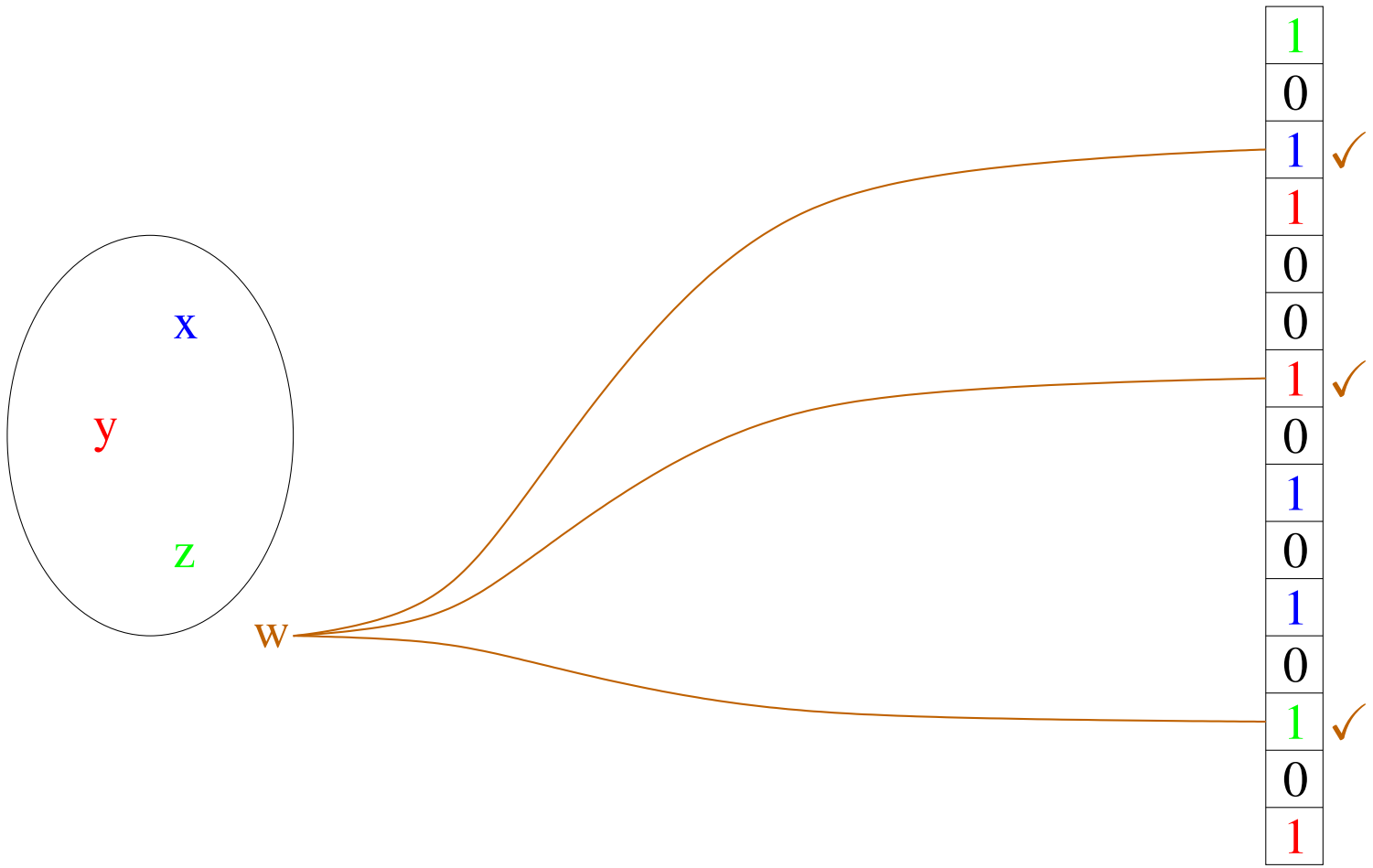


Insertion de l'élément z



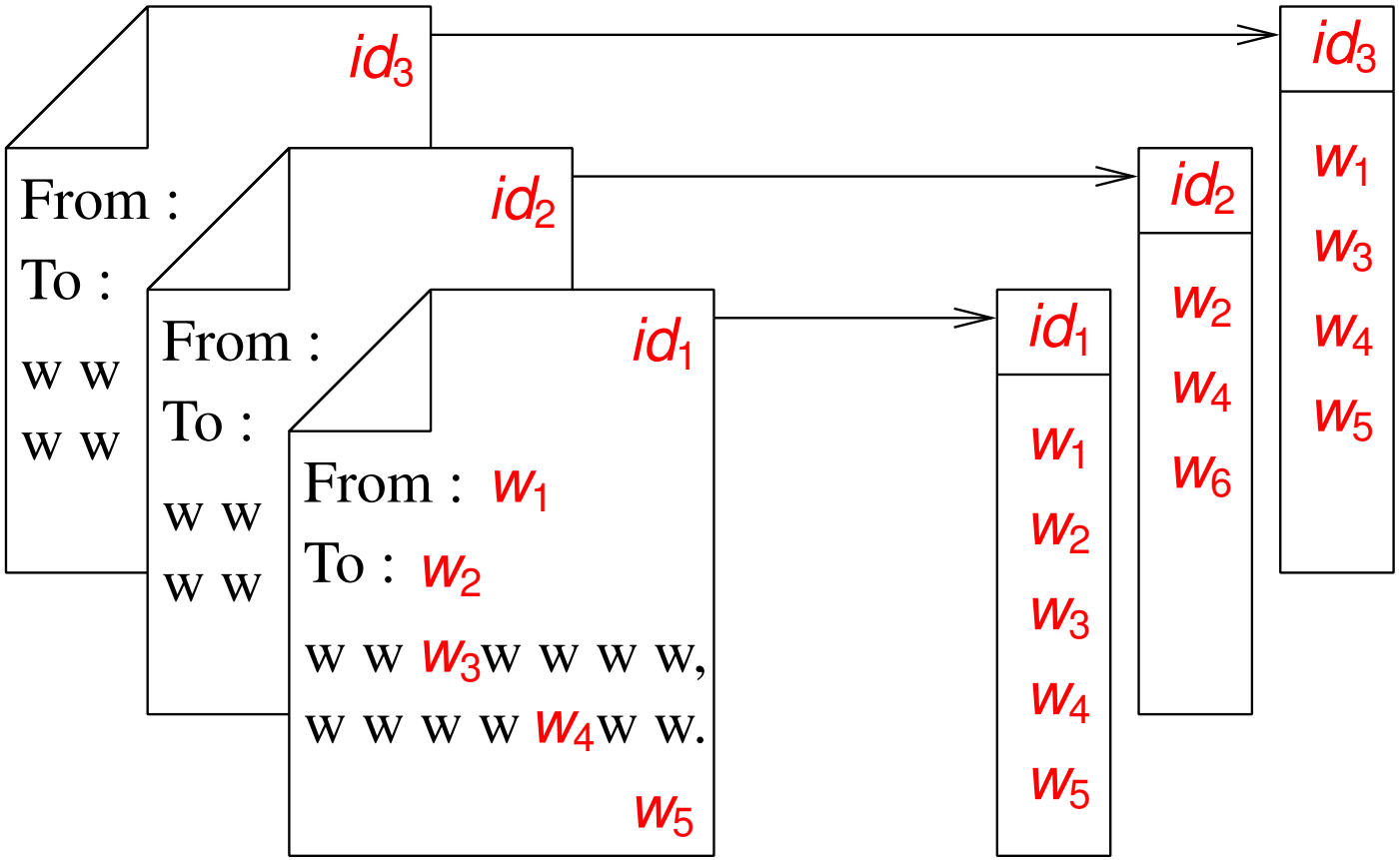
Faux positif

Test d'appartenance d'un élément $w \notin \{x, y, z\}$



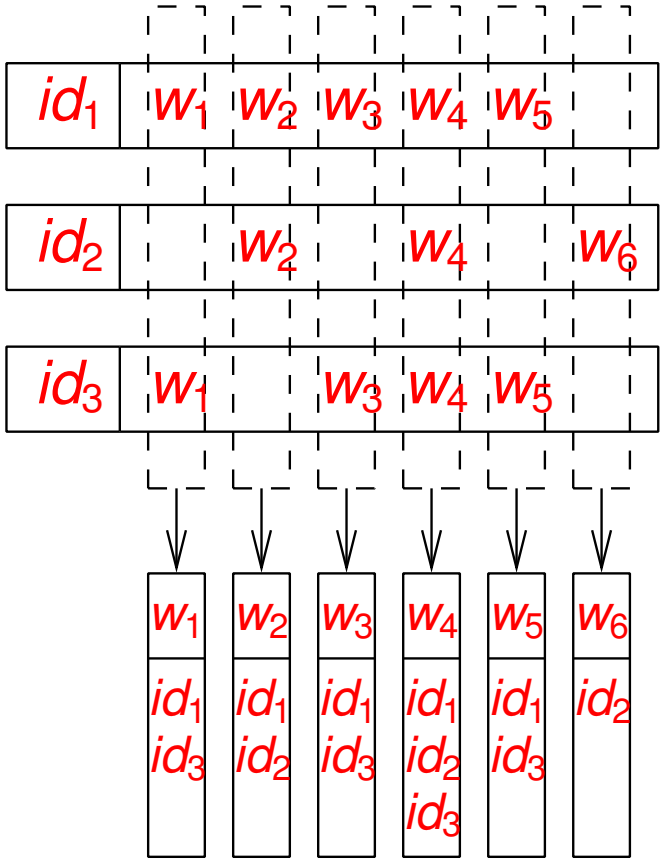
Plan

Indexation



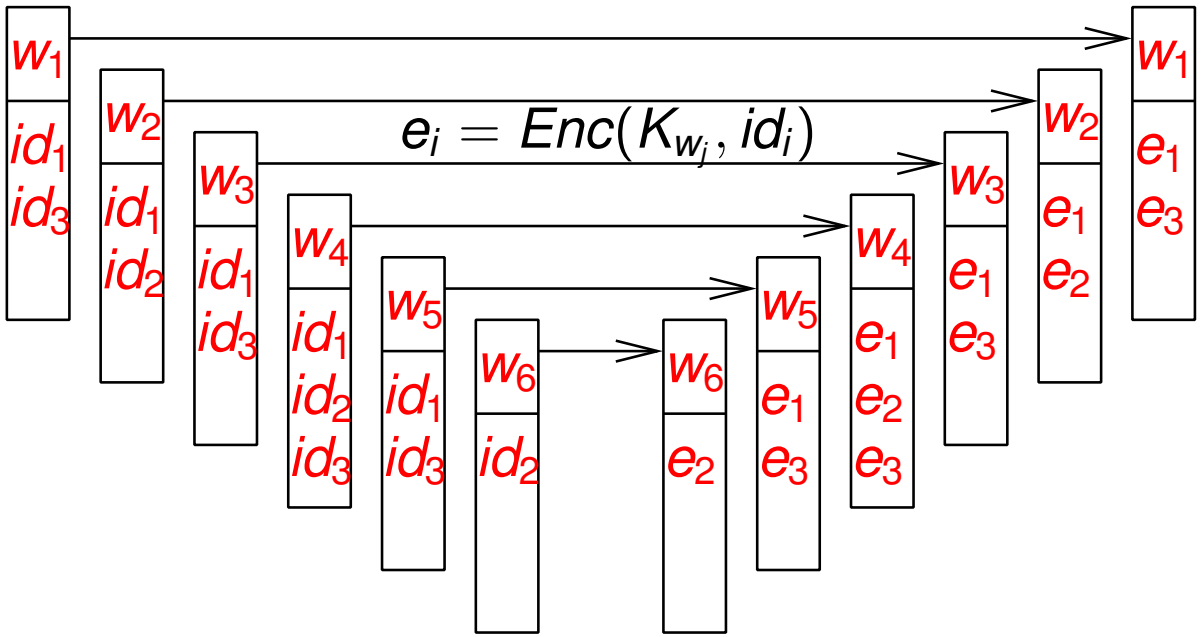
$|DB| = 3$ et $|W| = 6$

Indexation



$$|N| = 2 + 2 + 2 + 3 + 2 + 1 = 12$$

Chiffrement des identifiants

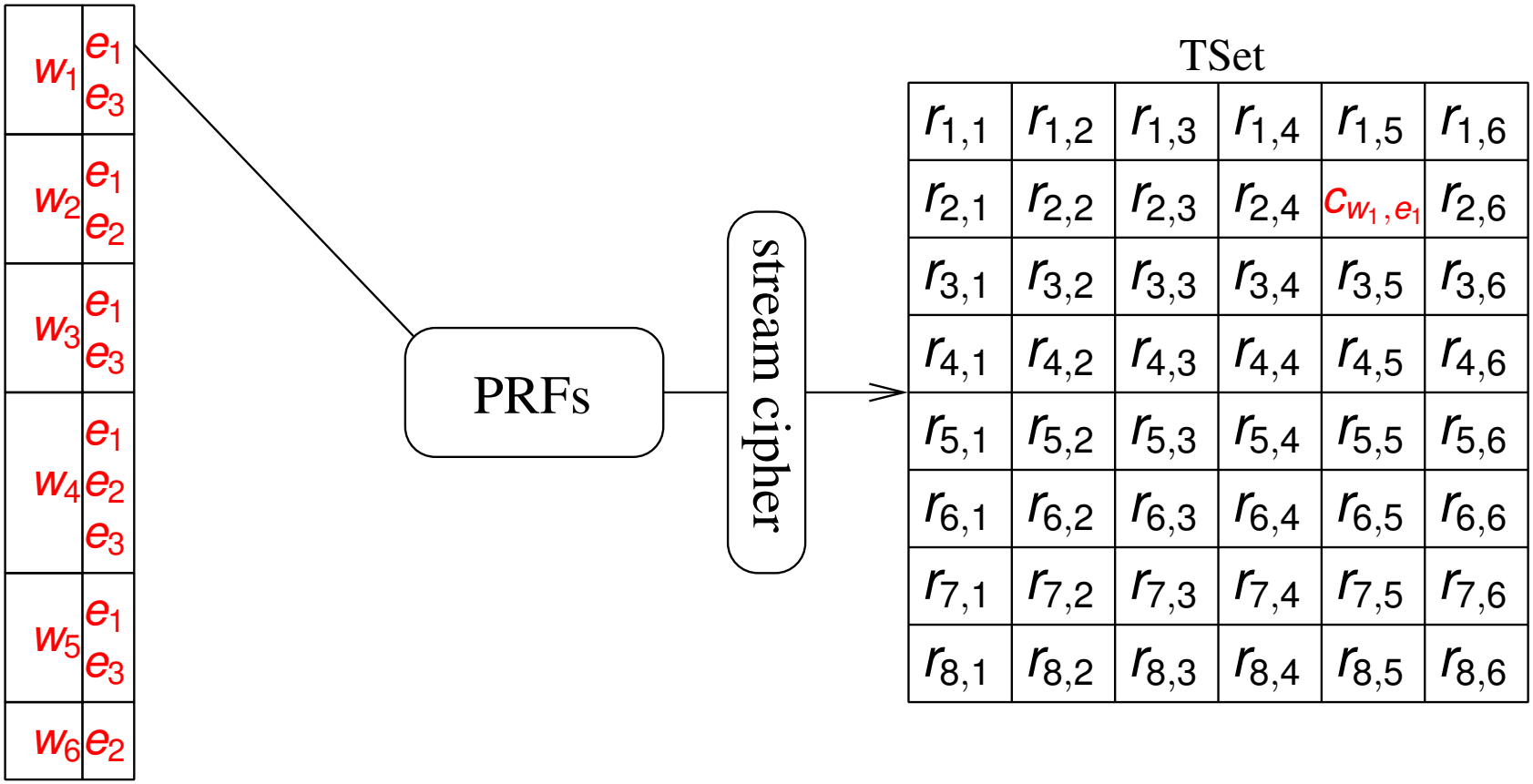


Création du répertoire de recherche

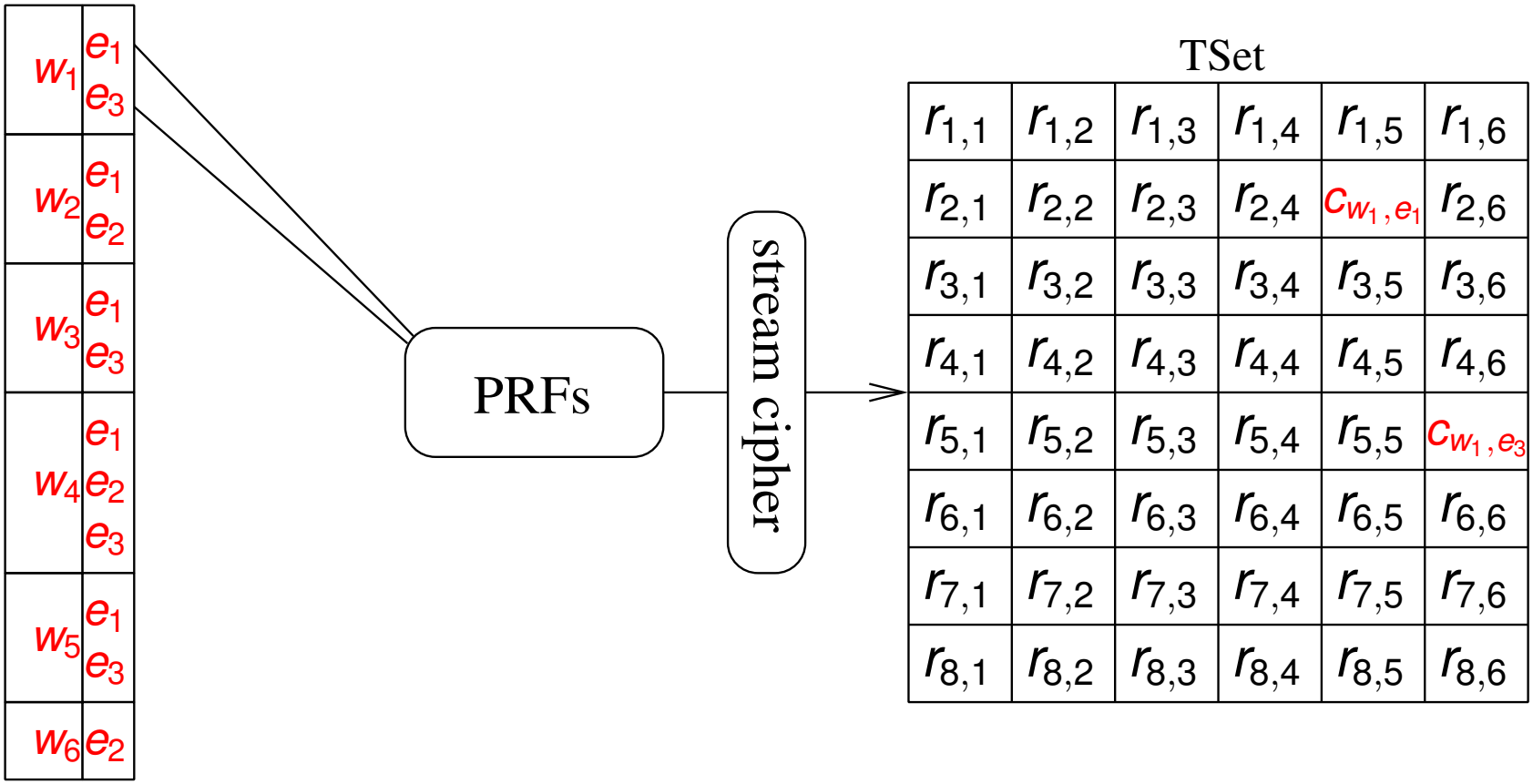
TSet

$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	$r_{1,4}$	$r_{1,5}$	$r_{1,6}$
$r_{2,1}$	$r_{2,2}$	$r_{2,3}$	$r_{2,4}$	$r_{2,5}$	$r_{2,6}$
$r_{3,1}$	$r_{3,2}$	$r_{3,3}$	$r_{3,4}$	$r_{3,5}$	$r_{3,6}$
$r_{4,1}$	$r_{4,2}$	$r_{4,3}$	$r_{4,4}$	$r_{4,5}$	$r_{4,6}$
$r_{5,1}$	$r_{5,2}$	$r_{5,3}$	$r_{5,4}$	$r_{5,5}$	$r_{5,6}$
$r_{6,1}$	$r_{6,2}$	$r_{6,3}$	$r_{6,4}$	$r_{6,5}$	$r_{6,6}$
$r_{7,1}$	$r_{7,2}$	$r_{7,3}$	$r_{7,4}$	$r_{7,5}$	$r_{7,6}$
$r_{8,1}$	$r_{8,2}$	$r_{8,3}$	$r_{8,4}$	$r_{8,5}$	$r_{8,6}$

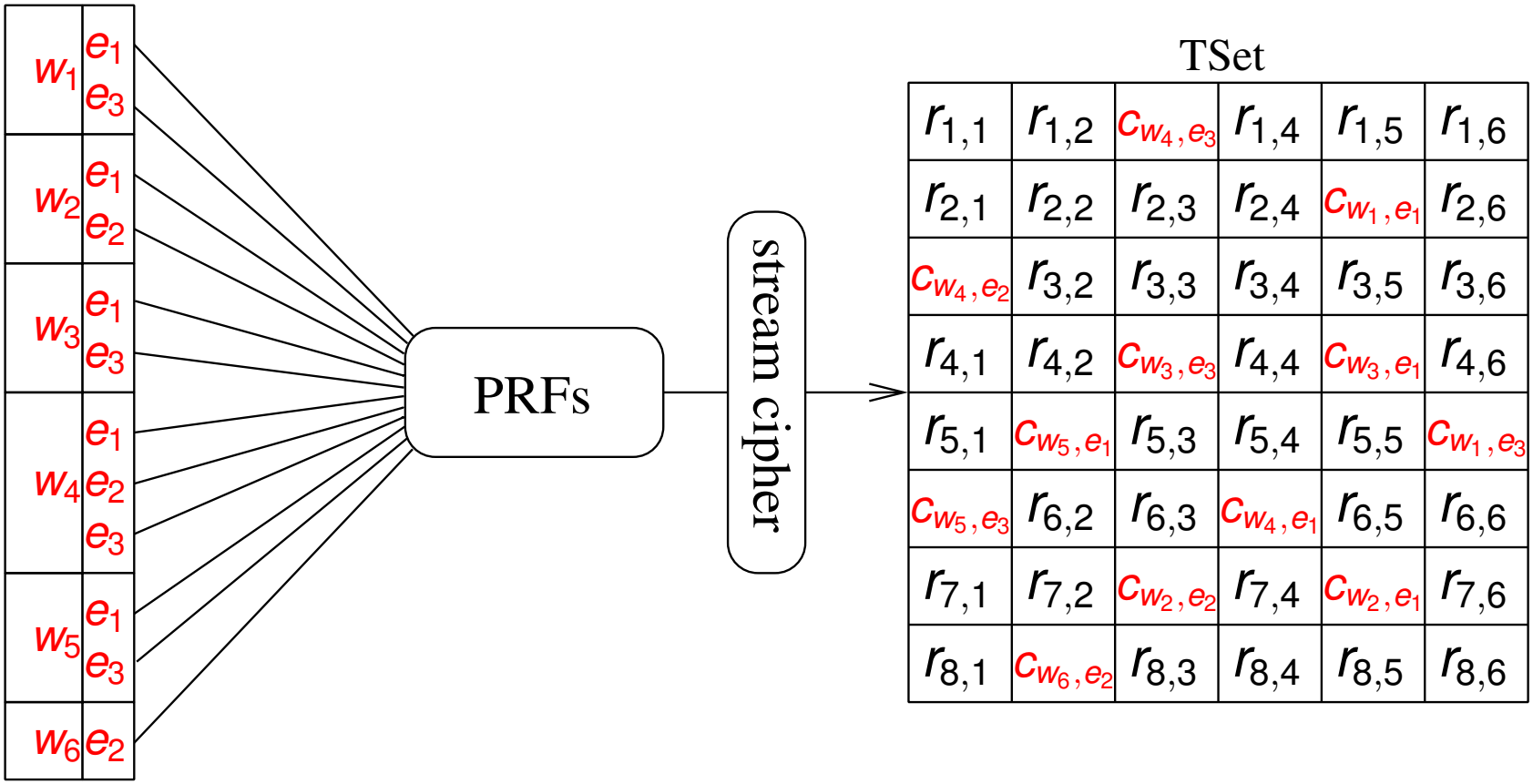
Création du répertoire de recherche



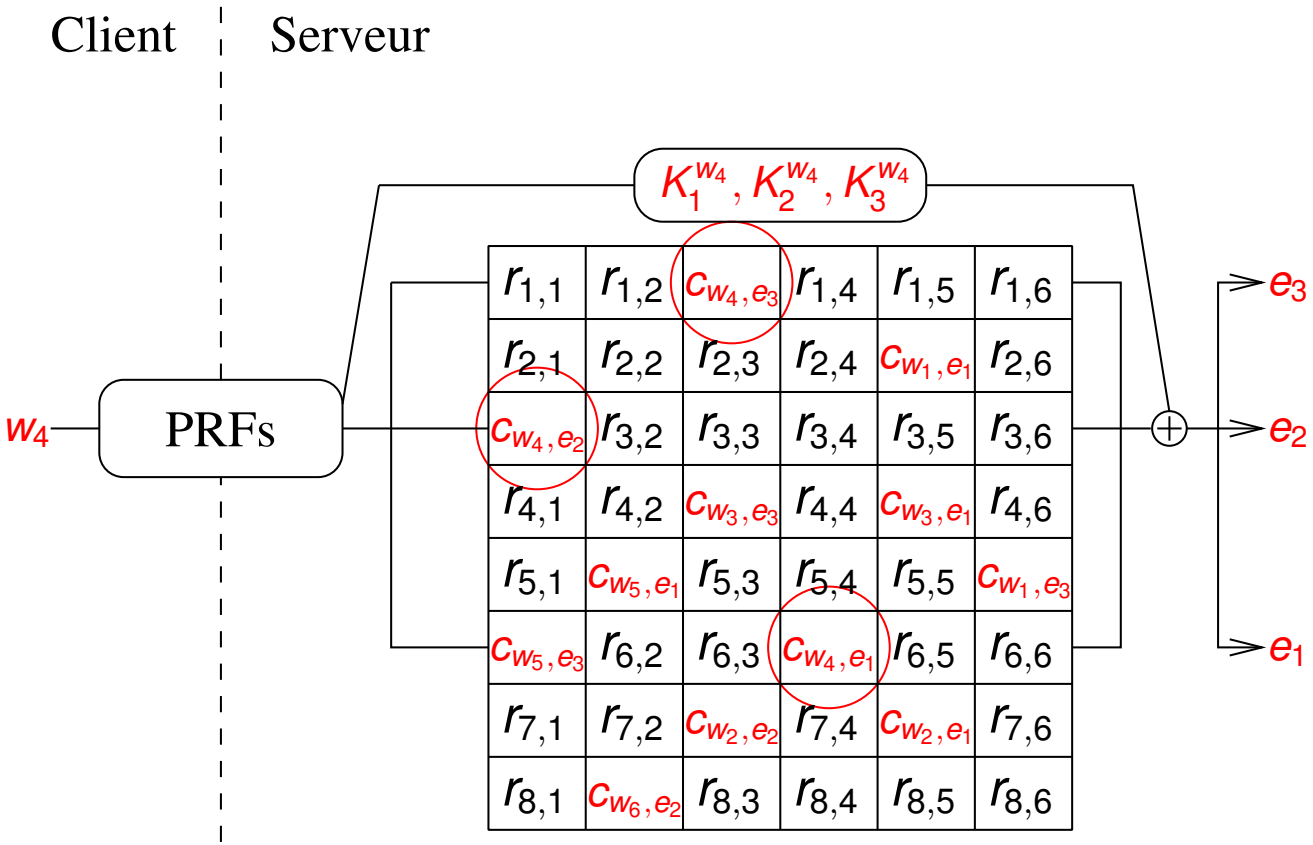
Création du répertoire de recherche



Création du répertoire de recherche



Recherche



$$id_i = Dec(K_{w_4}, e_i)$$

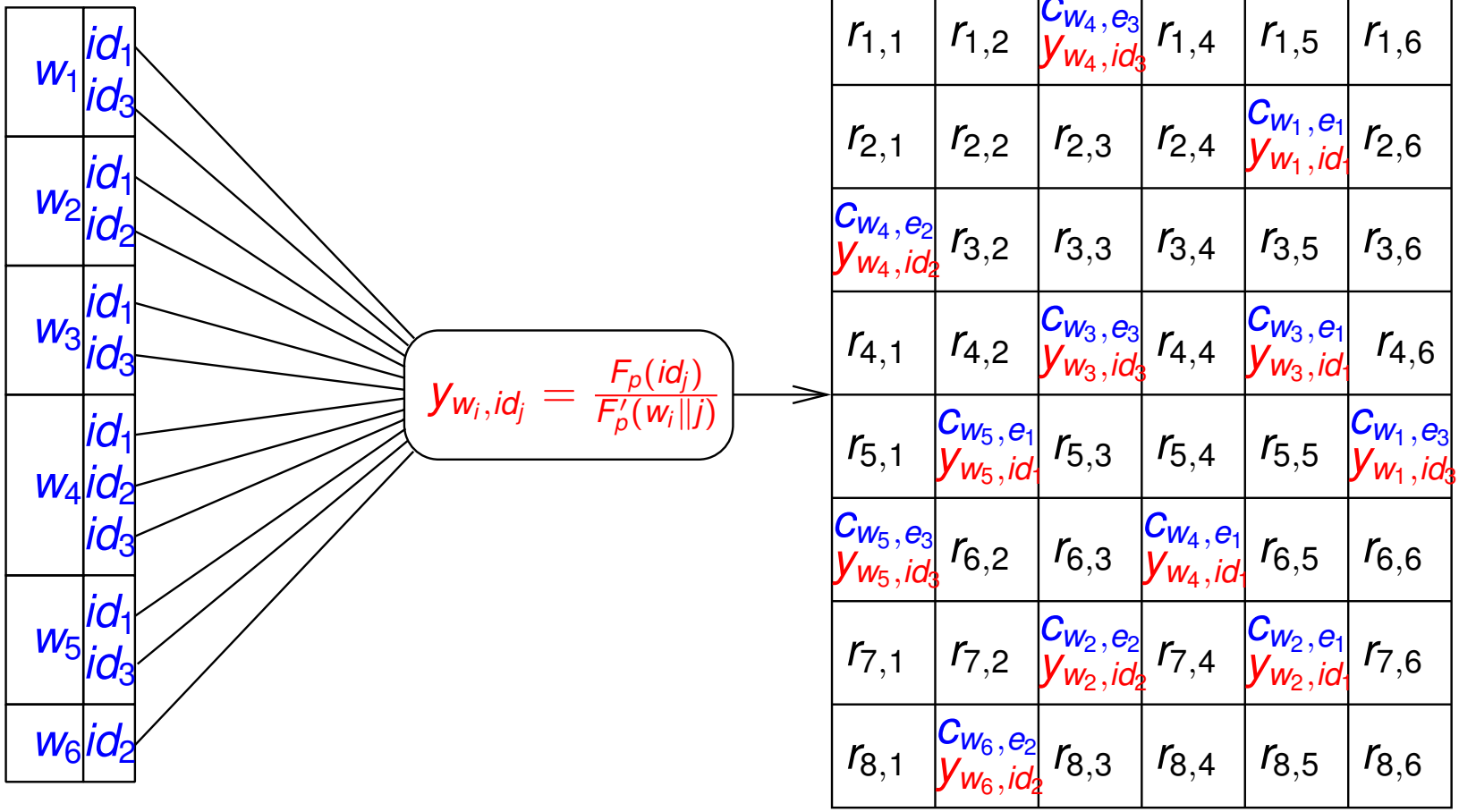
Résultat

Après déchiffrement le client obtient id_1, id_2 et id_3

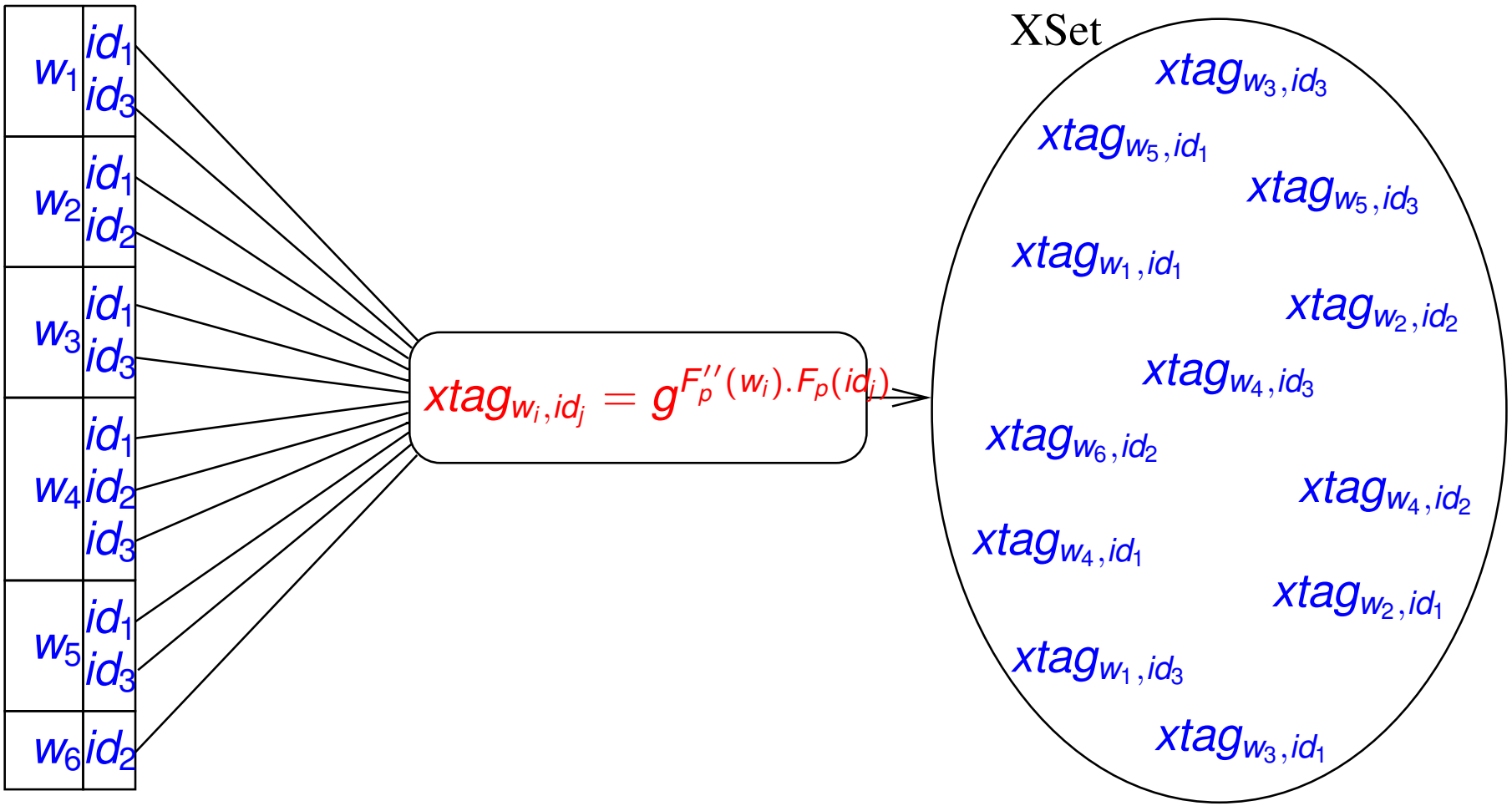
id_1	w_1 w_2 w_3 w_4 w_5
id_2	w_2 w_4 w_6
id_3	w_1 w_3 w_4 w_5

Plan

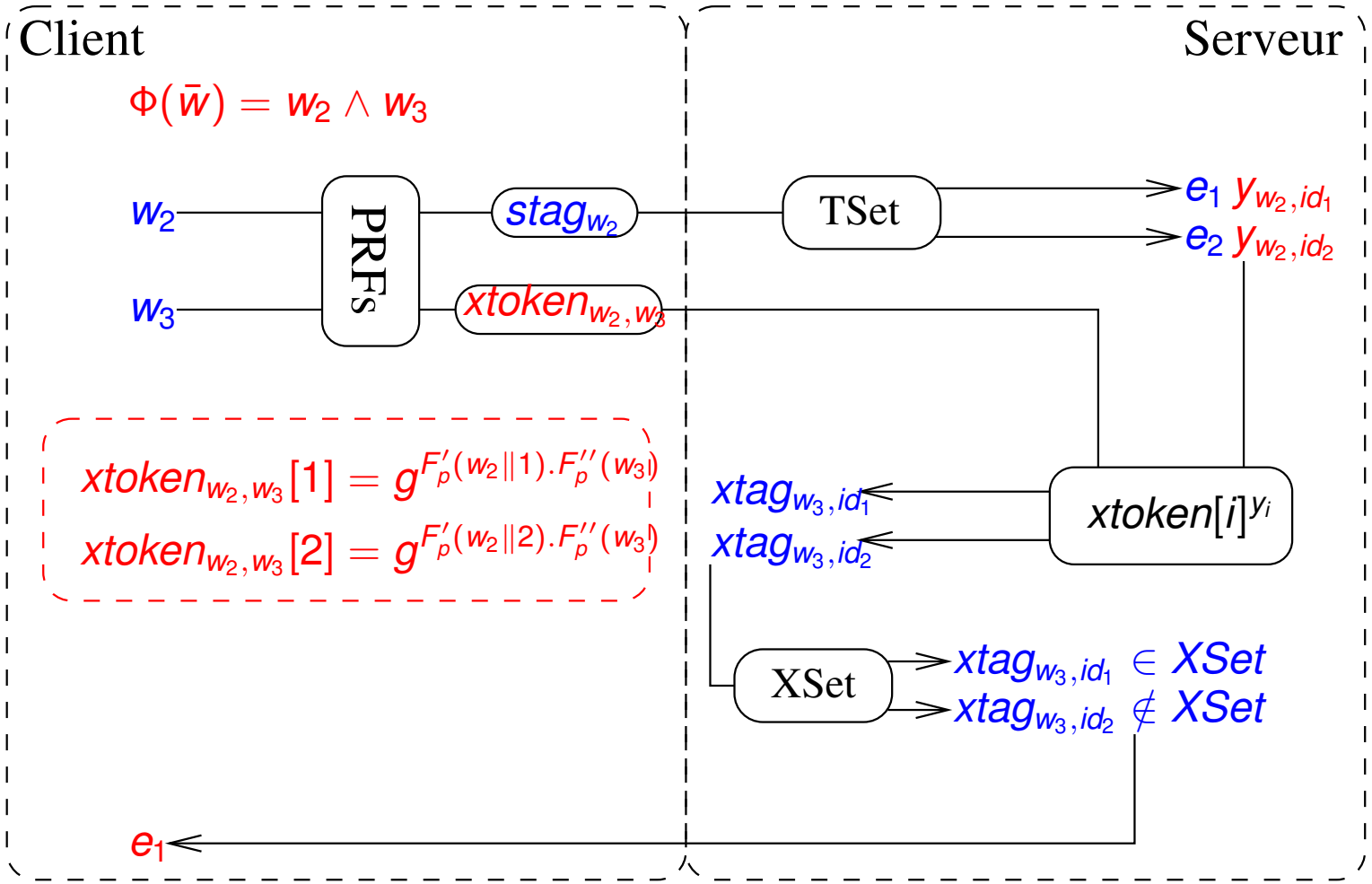
Modification du répertoire de recherche



Modification du filtre de Bloom



Recherche booléenne



Plan

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne

  ψ

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne
  ψ
- L'égalité des premiers mots-clés pour deux requêtes

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne
  ψ
- L'égalité des premiers mots-clés pour deux requêtes
  $w_1 \wedge w_2$ et $w_1 \wedge w_3$

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne
  ψ
- L'égalité des premiers mots-clés pour deux requêtes
  $w_1 \wedge w_2$ et $w_1 \wedge w_3$
- Le nombre de documents correspondant au premier mot-clé

Sécurité & Fonction de Fuite - 1

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuite liée à la base de données chiffrée

- N Le nombre de couples document/mot-clé

Fuites liées aux requêtes :

- La formule booléenne
  ψ
- L'égalité des premiers mots-clés pour deux requêtes
  $w_1 \wedge w_2$ et $w_1 \wedge w_3$
- Le nombre de documents correspondant au premier mot-clé
  $|DB(w_1)|$ si $\bar{w} = \{w_1, \dots\}$

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête $|\bar{w}|$

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête $|\bar{w}|$
- Le nombre de documents correspondant à une requête

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête $|\bar{w}|$
- Le nombre de documents correspondant à une requête $|DB(\psi(\bar{w}))|$

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête $|\bar{w}|$
- Le nombre de documents correspondant à une requête $|DB(\psi(\bar{w}))|$
- Les identifiants chiffrés correspondant à deux premiers mots-clés,
  pour deux requêtes identiques à l'exception de ceux-ci

Sécurité & Fonction de Fuite - 2

Le schéma [Cash et al., 2014] est prouvé sûr pour la fonction de fuite suivante :

Fuites liées aux requêtes :

- Le nombre de mots-clés distincts de chaque requête $|\bar{w}|$
- Le nombre de documents correspondant à une requête $|DB(\psi(\bar{w}))|$
- Les identifiants chiffrés correspondant à deux premiers mots-clés,
  pour deux requêtes identiques à l'exception de ceux-ci
  $DB(w_1 \wedge w_2)$ si $\bar{w}_1 = w_1||\bar{w}$ et $\bar{w}_2 = w_2||\bar{w}$

Plan

Ensemble documentaire & Matériel

Documents

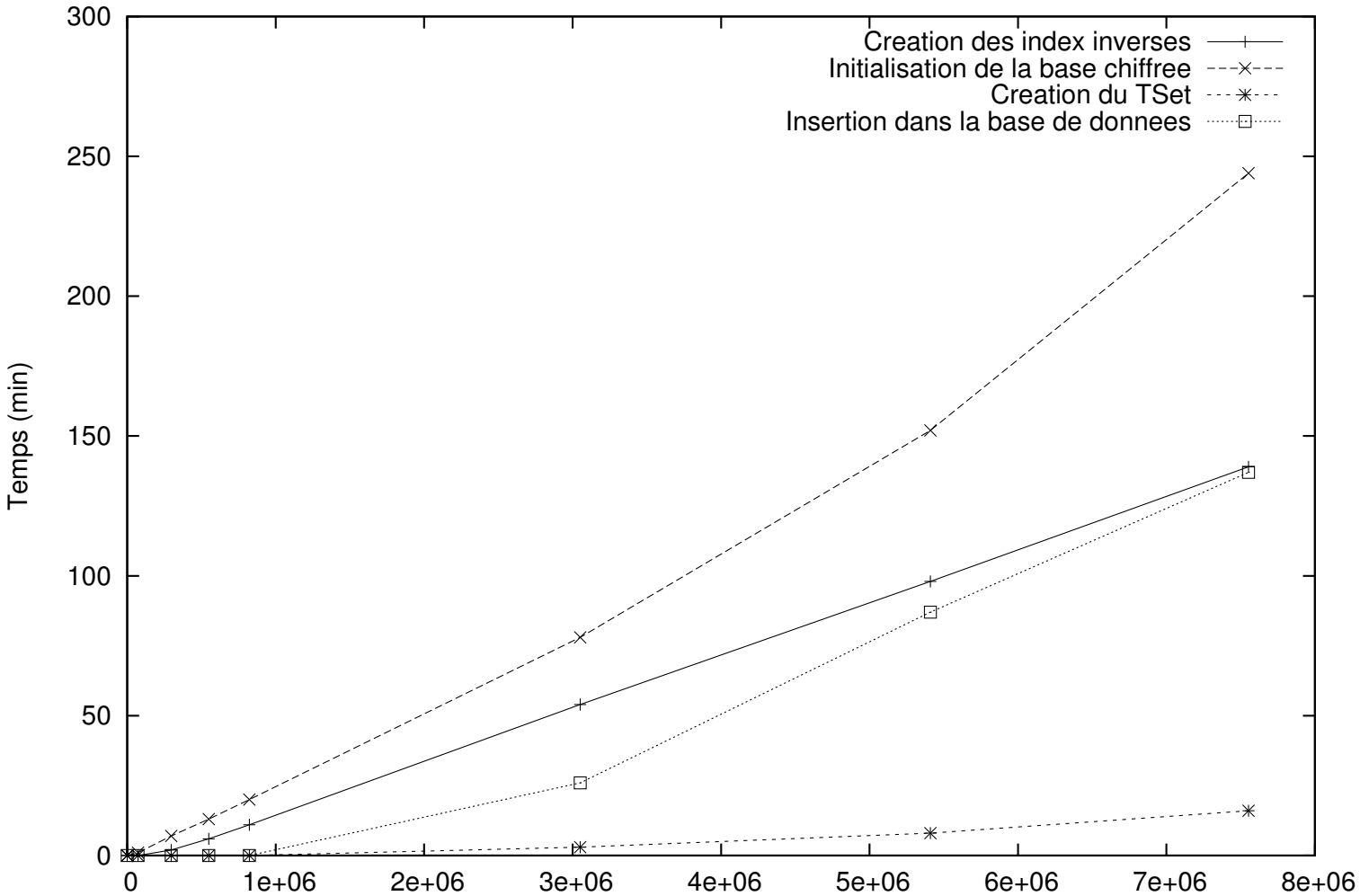
Corpus de courriers électroniques Enron : 600 000 courriers électroniques générés par 158 salariés

Matériel

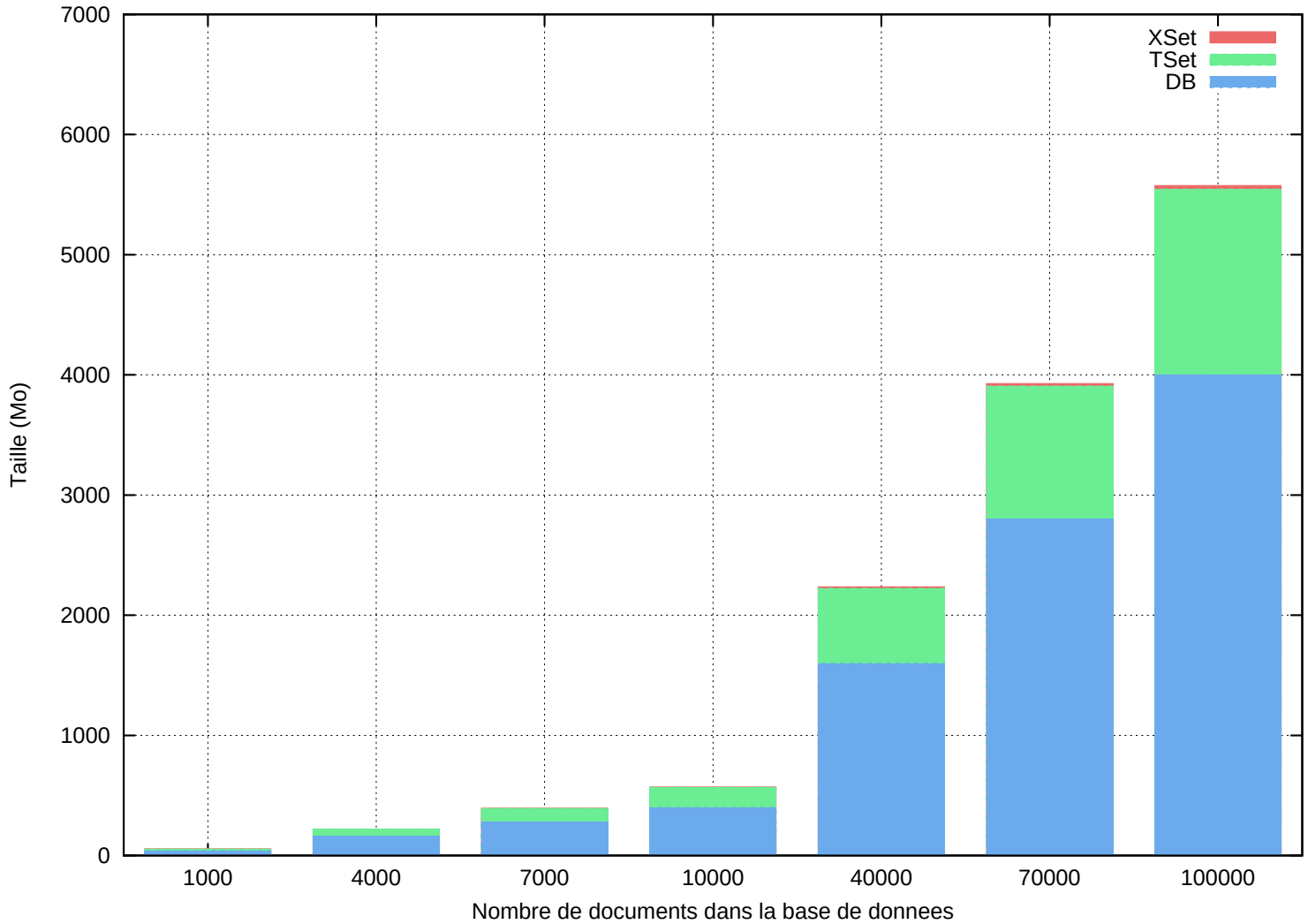
- *Intel quad-core* cadencé à $3,2$ Ghz
- 4 Go de RAM
- Disque dur de 7 200 RPM
- Système d'exploitation *Debian* 32 bits
- Base de données *PostgreSQL*

Plan

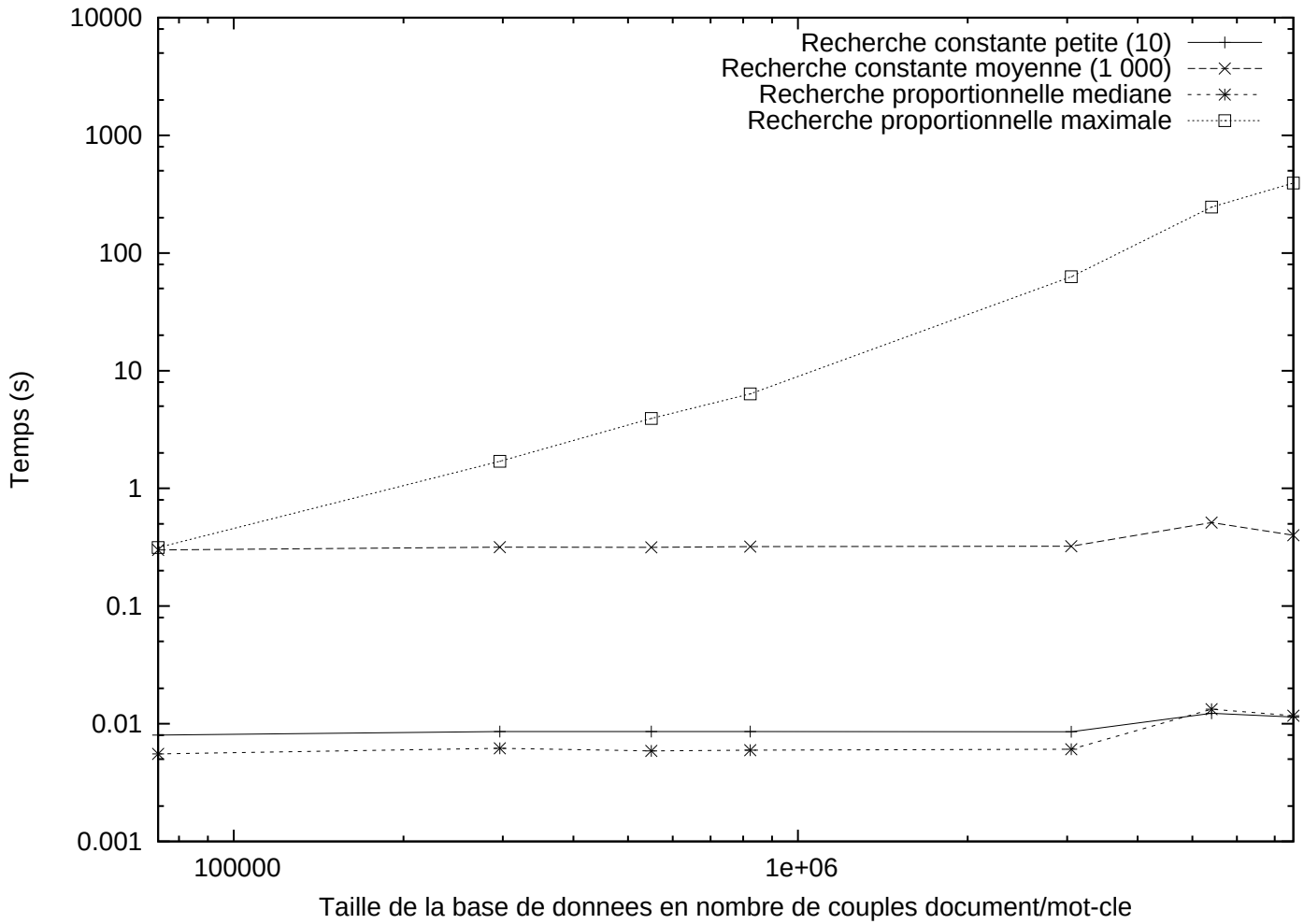
Temps d'initialisation



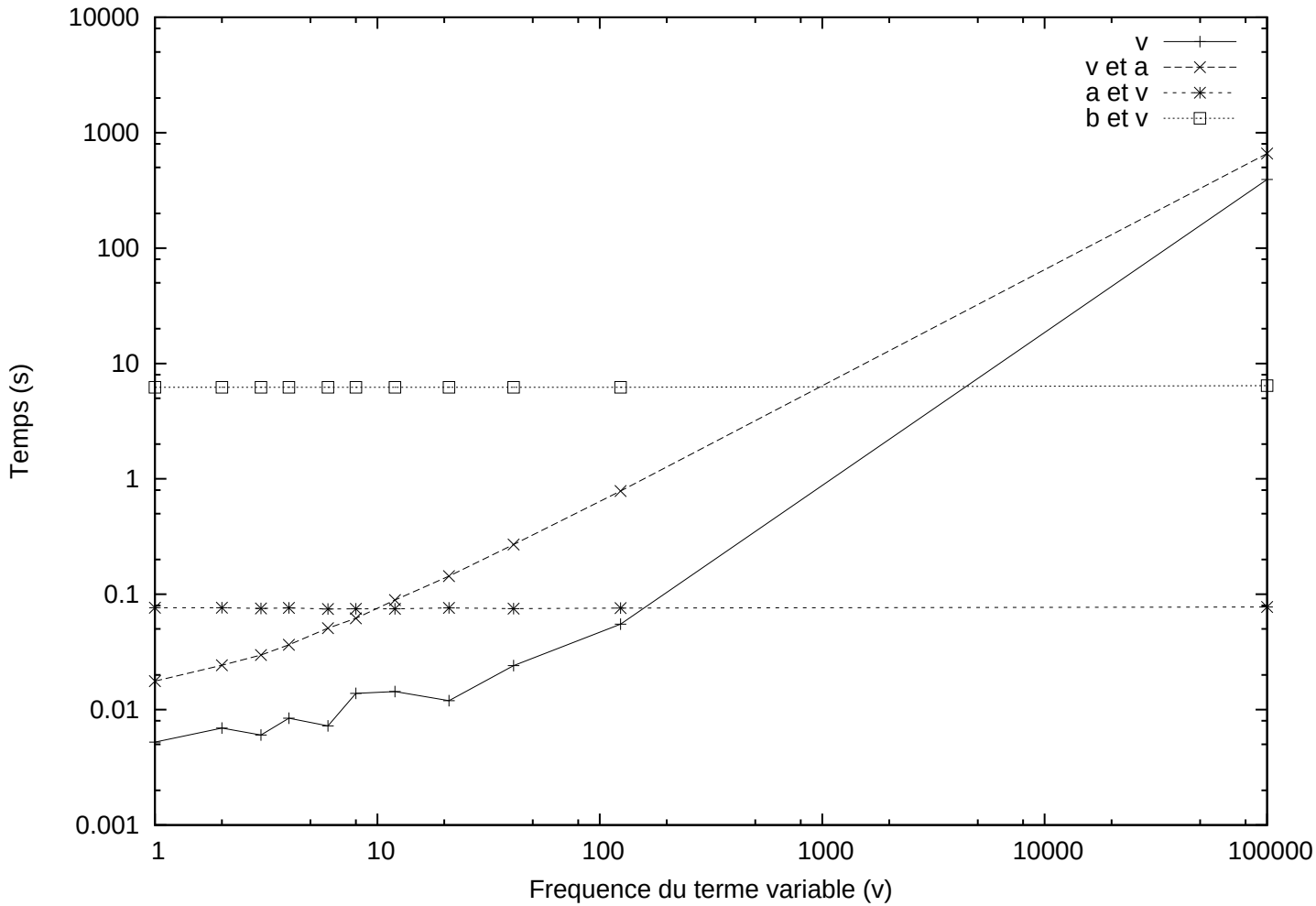
Taille de la base de données chiffrée



Recherche par mot-clé unique



Recherche booléenne



Conclusion

Nous avons réalisé un démonstrateur de recherche sur grands volumes de données chiffrées

Conclusion

Nous avons réalisé un démonstrateur de recherche sur grands volumes de données chiffrées

Sécurité ✓

Garantissant que l'opérateur de Cloud n'apprenne ni le contenu des requêtes ni le contenu des réponses via une preuve de sécurité

Conclusion

Nous avons réalisé un démonstrateur de recherche sur grands volumes de données chiffrées

Sécurité ✓

Garantissant que l'opérateur de Cloud n'apprenne ni le contenu des requêtes ni le contenu des réponses via une preuve de sécurité

Fonctionnalité ✓

Offrant à l'utilisateur un service de recherche évolué par mots-clés

Conclusion

Nous avons réalisé un démonstrateur de recherche sur grands volumes de données chiffrées

Sécurité ✓

Garantissant que l'opérateur de Cloud n'apprenne ni le contenu des requêtes ni le contenu des réponses via une preuve de sécurité

Fonctionnalité ✓

Offrant à l'utilisateur un service de recherche évolué par mots-clés

Efficacité ✓

Assurant un niveau de performances comparable à celui d'une solution non sécurisée

- Migration d'une grande base de données en moins de 48 h ✓
- Recherche transparente du point de vue de l'utilisateur ✓

Perspectives - Performances

Accès disque

- Utiliser un disque à mémoire flash SSD ;
- Stocker les structures de recherche en RAM.

Multiplication scalaire sur les courbes elliptiques

- Pré-calculs d'une table de multiples du générateur ;
- Implémentation spécifique à une courbe (Edwards) et à une architecture.

Parallélisation

- Implémenter une parallélisation intra-requête ;
- Paralléliser l'initialisation côté client.

Perspectives - Sécurité

Gestions des droits d'accès

La gestion du droit d'accès peut être faite par la Crypto et cela réduit les fuites d'informations.

Récupération des documents chiffrés

Utilisation de primitives avancées ex : Private Information Retrieval.

Réduction des fuites par masquage

- Bourrage des structures de données ;
- Masquage algorithmique via la fonction d'indexation.

Merci !

Questions ?

Références Bibliographiques I


Bloom, B. (1970).
Space/time trade-offs in hash coding with allowable errors.
Communications of ACM, 13(7):422–426.


Cash, D., Jaeger, J., Jarecki, S., Jutla, C., Krawczyk, H., Rosu, M.-C., and Steiner, M. (2013).
Dynamic searchable encryption in very-large databases : Data structures and implementation.
IACR Cryptology ePrint Archive.


Cash, D., Jarecki, S., Jutla, C., Krawczyk, H., Rosu, M.-C., and Steiner, M. (2014).
Highly scalable searchable symmetric encryption with support for boolean queries.
IACR Cryptology ePrint Archive.