IA SUR LES DONNÉES CROMPRESSÉES

Equipe: Sokunthy SROEM, Lounes DOUAR, Mamadou KAMARA, Serigne Fallou FALL

Encadrants : Michèle WIGGER, Thomas STURMA

Références : Direct Analytics of Generalized Deduplication Compressed IoT Data,aron Hurst, Qi Zhang and Daniel E. Lucani DIGIT,

PROBLÈME

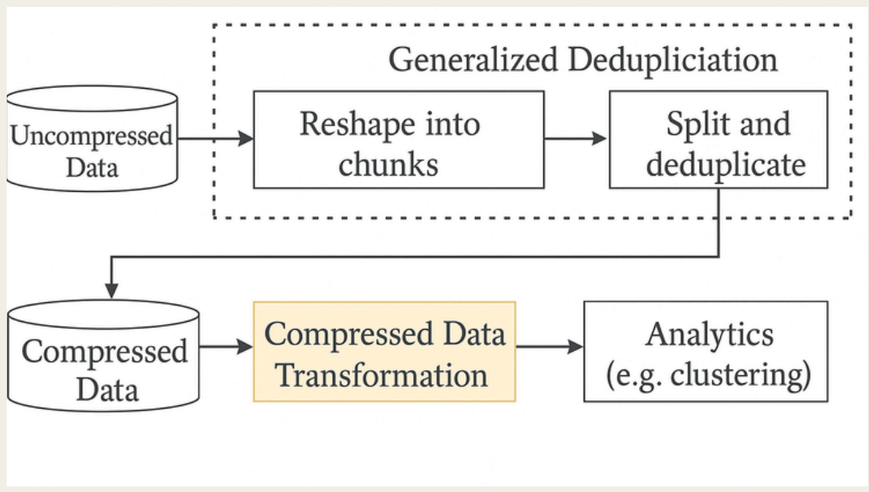
Les dispositifs IoT produisent des volumes massifs de données, créant des défis majeurs liés au stockage, à la transmission et à l'analyse. Un problème notable est le goulot d'étranglement dû à la nécessité de décompresser ces données avant de les analyser, entraînant des coûts élevés en termes de performance et d'espace de stockage.

OBJECTIF : Développer une méthode permettant de réaliser directement du clustering (K-means, X-means) sur des données compressées, via la méthode Generalized Deduplication (GDD), afin de :

- Réduire considérablement l'espace de stockage.
- Maintenir la performance analytique.

MÉTHODOLOGIE:

1. Génération et Prétraitement des données :
   - Génération de jeux de données IoT simulés.
   - Application de la méthode GDD pour compresser ces données.
2. Application des méthodes de clustering :
   - K-means (détermination initiale des clusters).
   - X-means (optimisation automatique du nombre de clusters).



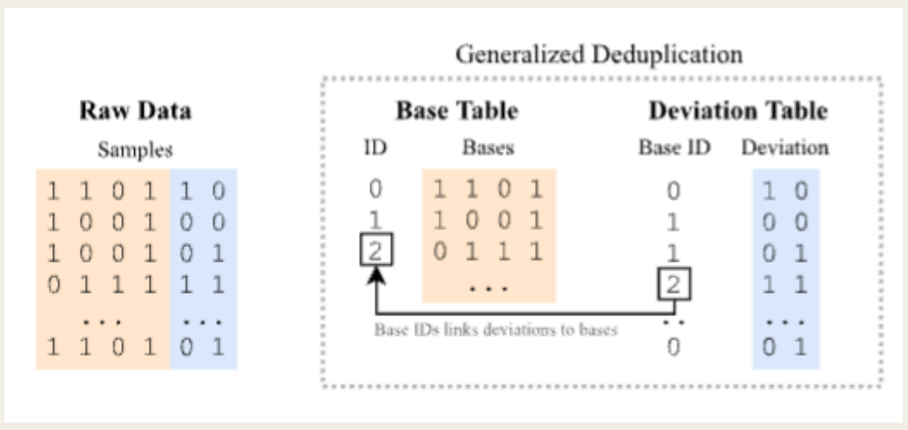
ALGORITHME DE DÉDUPLICATION GÉNÉRALISÉE

POURQUOI : Exploite cette structure en regroupant les morceaux de données similaires, permettant à la fois la compression et la préservation de la structure.

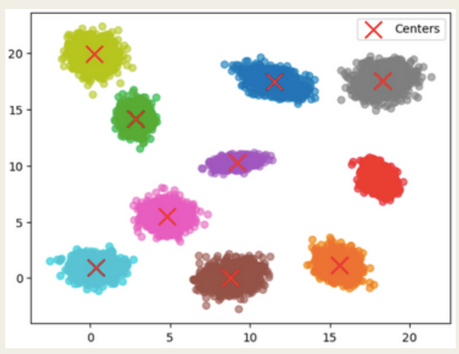
COMMENT: Chaque morceau de données est divisé en :

- Base : bits stables (très redondants)
- Déviation : bits variables (stockés tels quels)

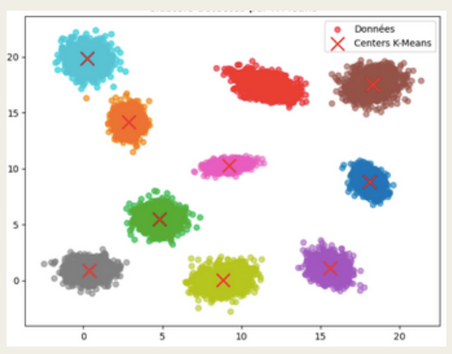
Les morceaux partageant la même base sont regroupés.



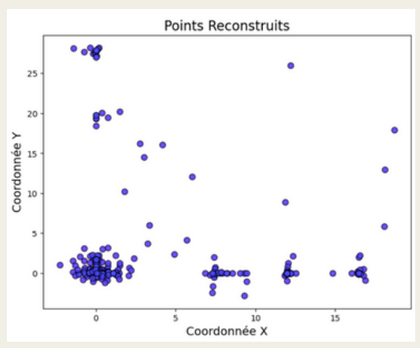
Génération de données



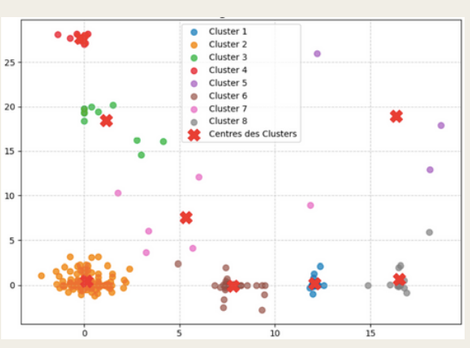
Clustering sur ces données générées

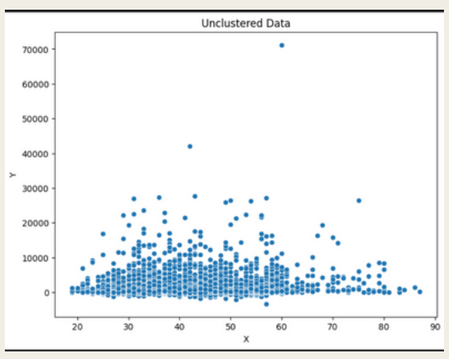


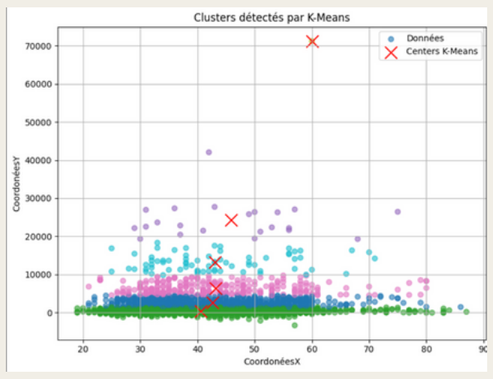
Compression (GDD- 9 premiers bits)

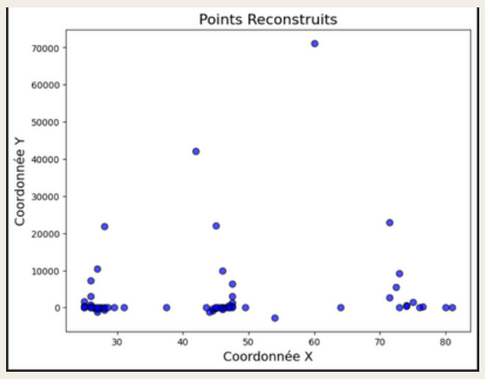


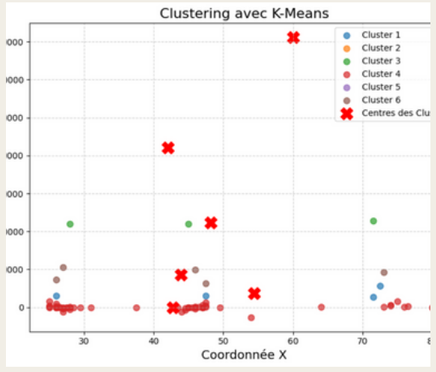
Clustering sur données compressées











PERFORMANCES SUR DES DONNÉES RÉELLES

Mesure	Valeur
Nb. Points originaux	4521
Nb. Bases retenues	70
Taille des bases (bits)	18 (9+9)
Taux de compression	0.832

TABLE 1 – Résultats quantitatifs - données bancaires

Indicateur	Non compressé	Compressé
Score silhouette	0.69	0.78
Temps K-means (s)	0.066	0.0091

TABLE 2 – Comparaison de clustering avant/après compression

CONCLUSION | BILAN

- L'approche utilisant GDD permet de réaliser efficacement du clustering directement sur des données compressées, réduisant le stockage des données et le temps pour traiter les données sans compromettre la qualité analytique.