Fouille de flots de données multidimensionnelles

Yoann Pitarch

Sous la direction d'**Anne Laurent**, **Marc Plantevit** et **Pascal Poncelet**

ARC SéSur INRIA

10 septembre 2008

Plan

De plus en plus de données

Ces dernières années, le volume des données a explosé.

- 166 millions de visiteurs par jour sur le site de Yahoo,
- 35 millions de compteurs communicants déployés par EDF d'ici à 2015,
- 30 milliards d'email envoyés par jour,
- Développement de la technologie RFID,
- ...

Besoin de méthodes d'analyse et d'interrogation adaptées à une telle quantité.

Pourquoi analyser de telles données ?

Enjeux commerciaux et financiers considérables !

- Enregistrement des appels téléphoniques,
- Analyse de la consommation électrique,
- Suivi des transactions par cartes banquaires,
- Supervision de réseau,
- Optimisation du processus industriel,
- Hopitaux,
- Clickstreams (analyse des pages visitées par les internautes),
- ...

L'analyse et l'interrogation d'un tel volume de données soulèvent des **difficultés** .

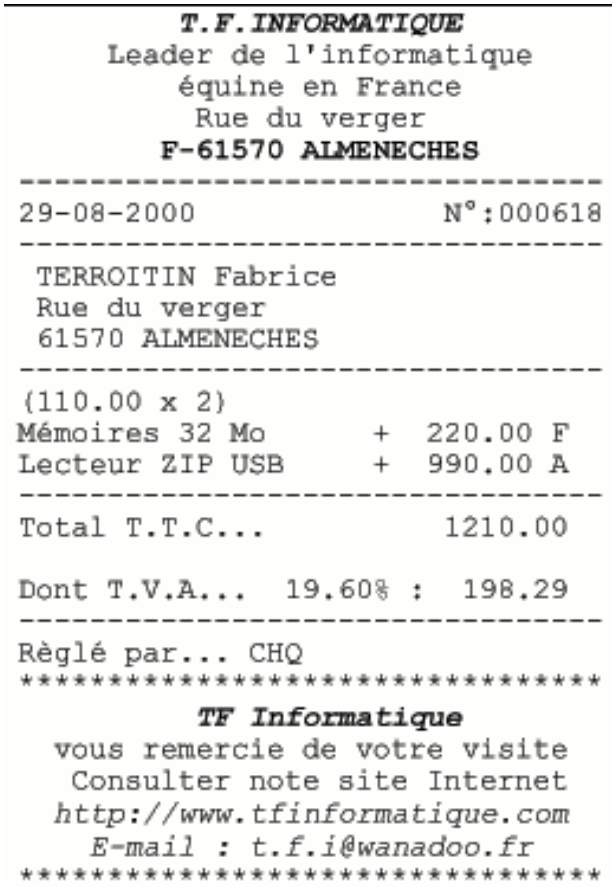
Les flots de données...

Flot : séquence **changeante** , (potentiellement) **infinie** de données **précises** et circulant **rapidement** .

- **Changeante** $\Rightarrow$ difficile de prédire les valeurs d'un flot
- **Infinie** $\Rightarrow$ créer une structure compacte pour résumer l'historique
- **Précises** $\Rightarrow$ peu intéressant pour un analyste (ventes par produits et par heure)
- **Rapidement** $\Rightarrow$ insertion rapide d'un élément dans la structure

Un flot $S{=}B_0, B_1, ..., B_n$ est une séquence infinie de batchs (B_n le plus récent). Un batch $B_i{=}\{\mathscr{T}_1, \mathscr{T}_2, \mathscr{T}_3, ..., \mathscr{T}_k\}$ est un ensemble de données qui arrive pendant les i^{iemes} unités de temps.

...multidimensionnelles



T.F.INFORMATIQUE
Leader de l'informatique
équine en France
Rue du verger
F-61570 ALMENECHES

29-08-2000 N°:000618

TERROITIN Fabrice
Rue du verger
61570 ALMENECHES

(110.00 x 2)
Mémoires 32 Mo + 220.00 F
Lecteur ZIP USB + 990.00 A

Total T.T.C... 1210.00

Dont T.V.A... 19.60% : 198.29

Règlé par... CHQ

****

TF Informatique
vous remercie de votre visite
Consulter note site Internet
http://www.tfinformatique.com
E-mail : t.f.i@wanadoo.fr

****

Multidimensionnelles

- produits,
- magasin,
- ...

Multi-niveaux

- produits, catégorie de produits
- magasin, département, région, ...

Requête

Lister les ventes de l'Hérault par catégories de ces deux dernier mois.

Prise en compte de la multidimensionnalité des données dans un environnement statique

- Un problème étudié depuis quelques années dans les entrepôts de données.
- Un modèle logique couramment utilisé : le cube de données

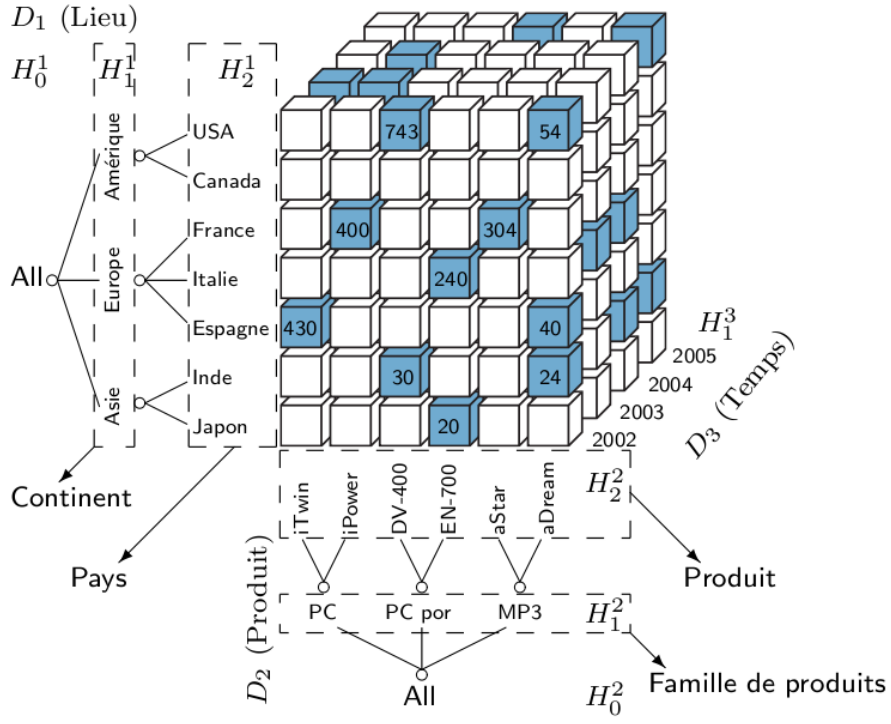


FIG.: Représentation graphique d'un cube de données [?]

Prise en compte de la multidimensionnalité des données dans un environnement statique

- Un problème etudié depuis quelques années dans les entrepôts de données.
- Un modèle logique couramment utilisé : le cube de données

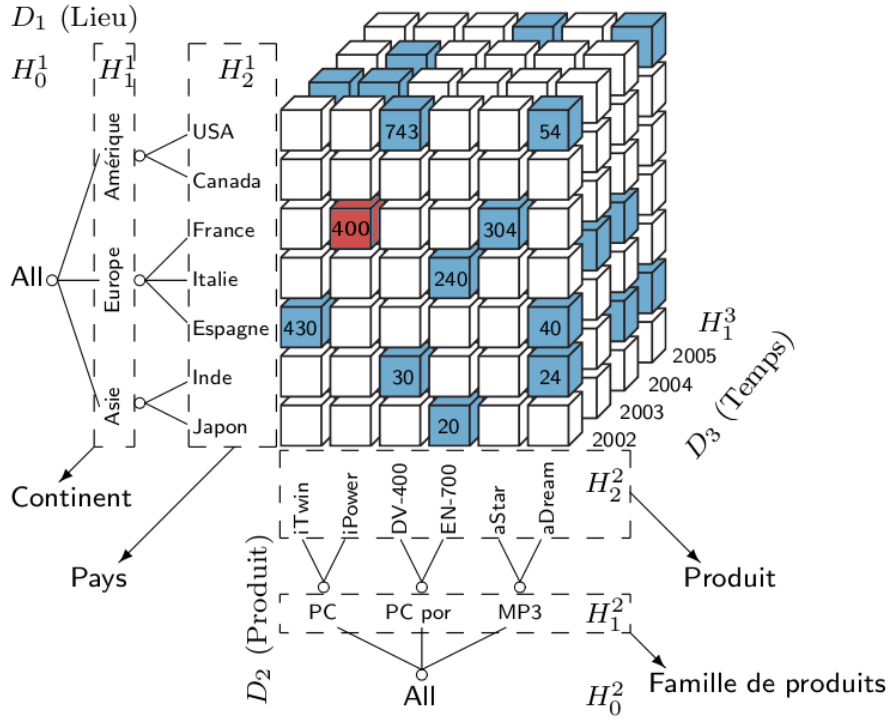


FIG.: Représentation graphique d'un cube de données [?]

Prise en compte de la multidimensionnalité des données dans un environnement statique

- Un problème étudié depuis quelques années dans les entrepôts de données.
- Un modèle logique couramment utilisé : le cube de données

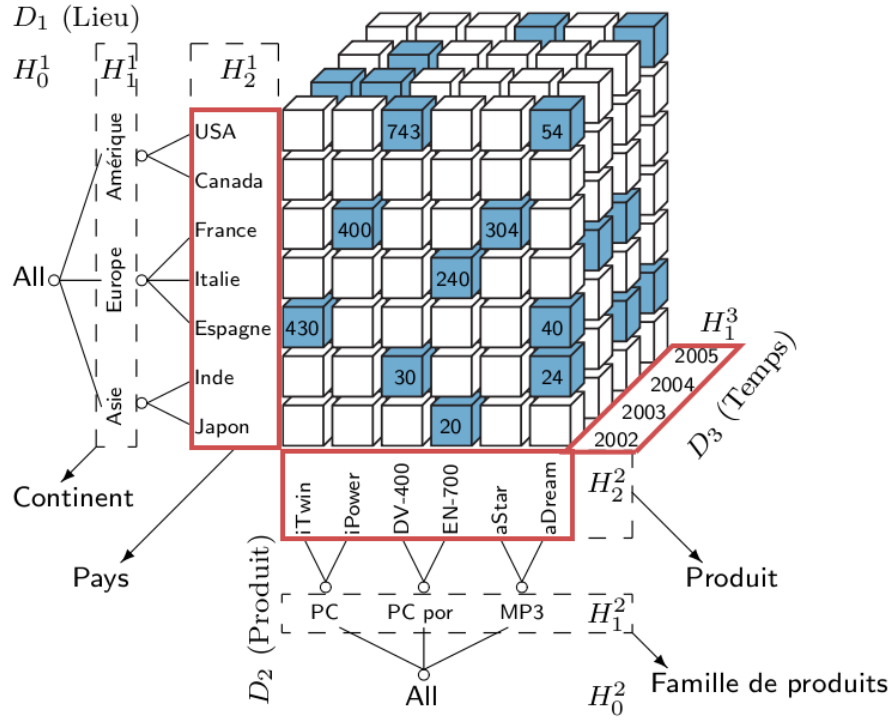


FIG.: Représentation graphique d'un cube de données [?]

Prise en compte de la multidimensionnalité des données dans un environnement statique

- Un problème étudié depuis quelques années dans les entrepôts de données.
- Un modèle logique couramment utilisé : le cube de données

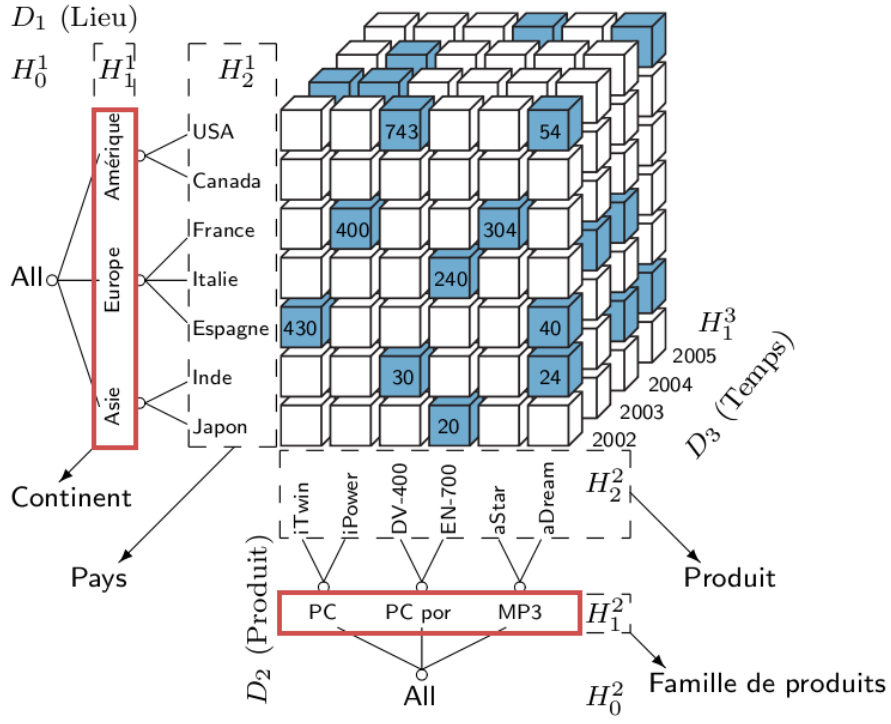


FIG.: Représentation graphique d'un cube de données [?]

Quels cuboïdes matérialiser ?

Une problématique présente dans un environnement statique...

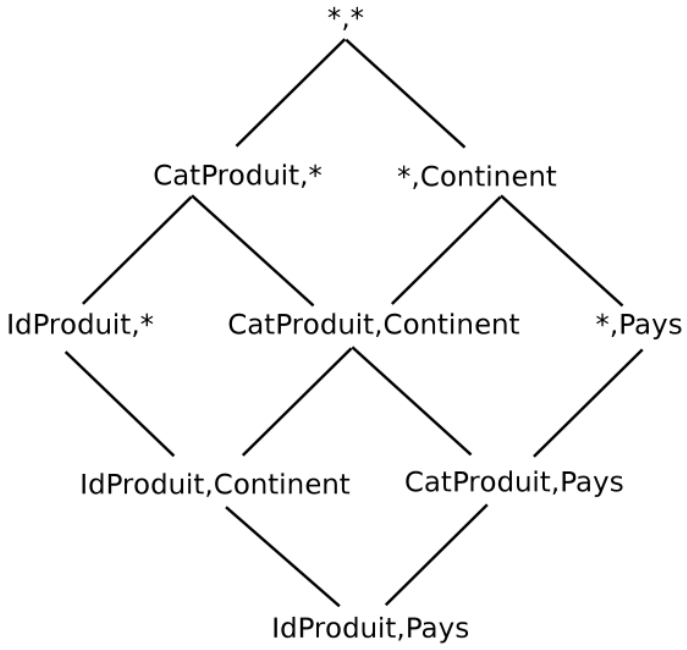


FIG.: Treillis des cuboïdes

...et qui est d'autant plus justifiée dans le contexte des flots de données.

Une vision *cube de données* est-elle réaliste dans un contexte de flots ?

Contraintes

Cube

- Vision mutlidimensionnelle
- Gestion des hiérarchies
- Structure volumineuse

Flot

- Granularité fine
- Débit rapide
- Taille infinie
- Une seule lecture des données

Problématique

Trouver une méthode de résumé de flots de données multidimensionnelles qui :

1. Minimise espace occupé
2. Maximise la qualité des réponses aux requêtes multi-niveaux et historiques

Plan

StreamCube [?]

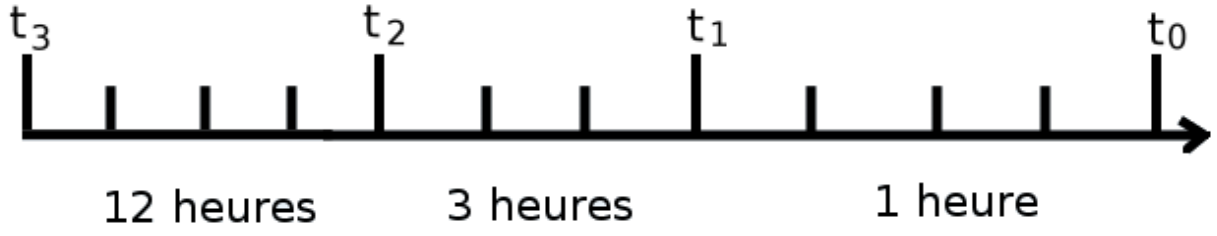
Une seule approche propose de gérer la multidimensionnalité des données d'un flot : **StreamCube** [?].

Principes

1. Compression de la dimension temporelle grâce au modèle des **tilted-time windows**
2. Choix de deux cuboïdes particuliers : les **critical layers**
3. Propagation dans le treillis en suivant le **popular path**

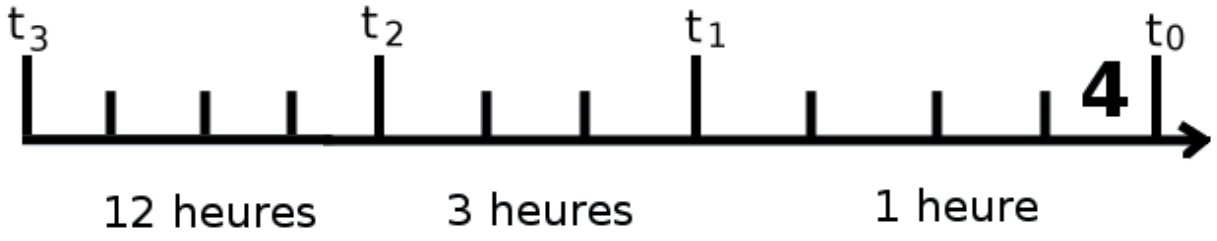
Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



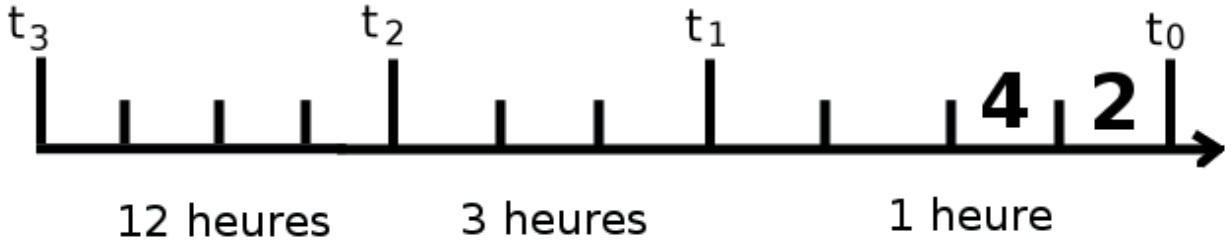
Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



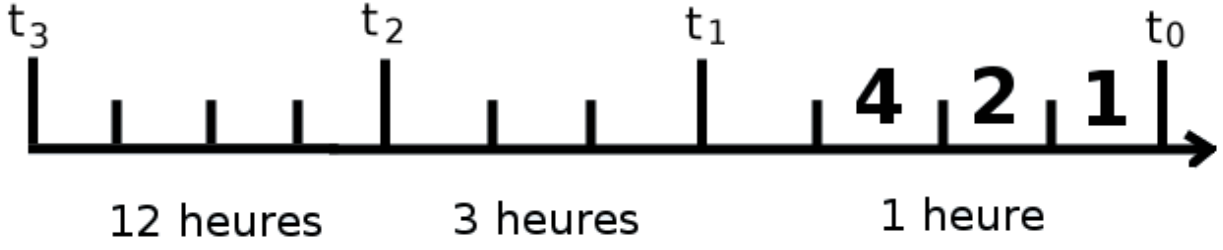
Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



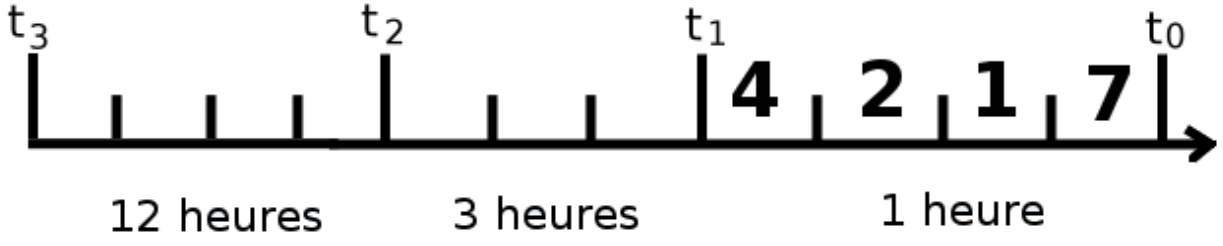
Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



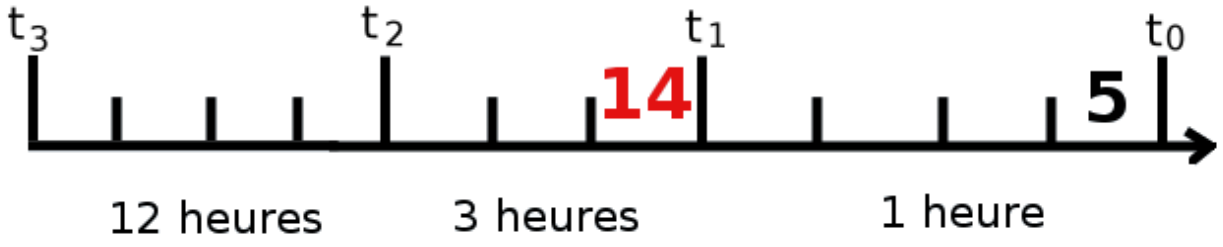
Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



Tilted-time window

- Inspiré du fonctionnement de la mémoire humaine
- Correspond au besoin des décideurs
- Découpage du temps
- Les élements récents sont stockés précisement
- Plus on s'éloigne, plus cette précision diminue
- Chaque mesure du cube est une tilted-time window



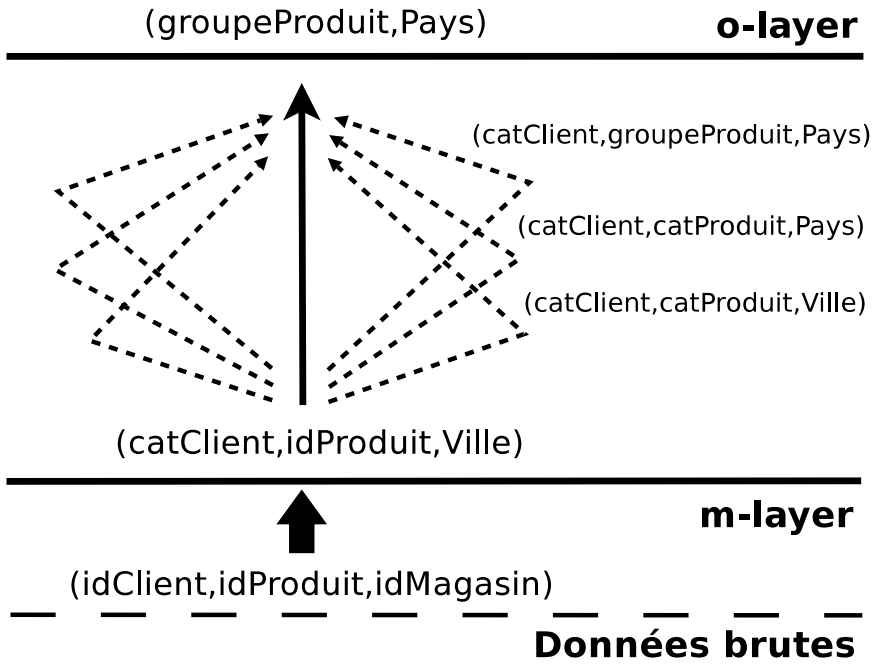
Critical layers et Popular path

Critical layers

- m-layer : le plus précis
- o-layer : le plus couramment consulté

Popular path

- Relie les critical layers
- Fixé par utilisateur



Les limites de StreamCube

Tout l'historique du flot est-il consulté sur tous les niveaux d'une hiérarchie ?

Par exemple, est-il intéressant de savoir qu'au cours de la dernière année, il y a eu 6548 oeufs vendus au Carrefour Trifontaine ?

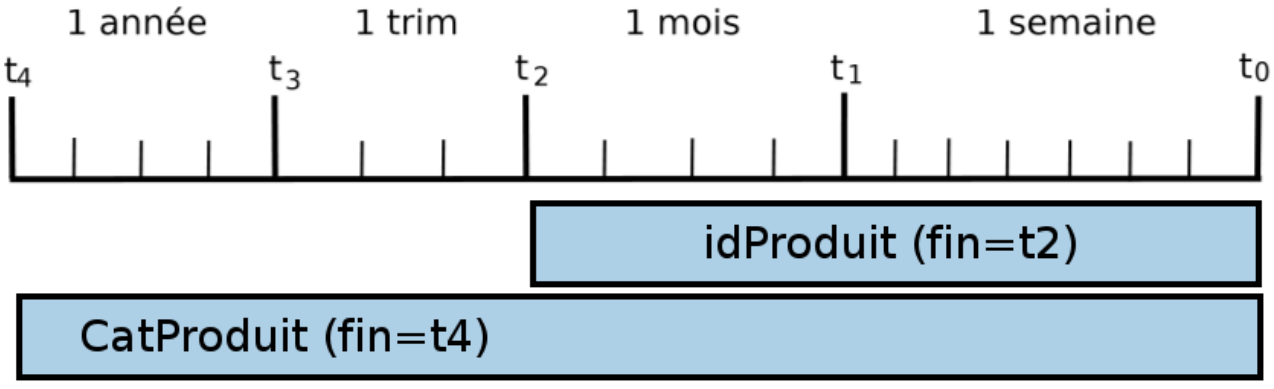
Est-il nécessaire de propager les valeurs du flot le long du popular path ?

Le cuboïde (catClient,catProduit,Ville) est calculable à partir du cuboïde (catClient,idProduit,Ville) grâce à la généralisation.

Plan

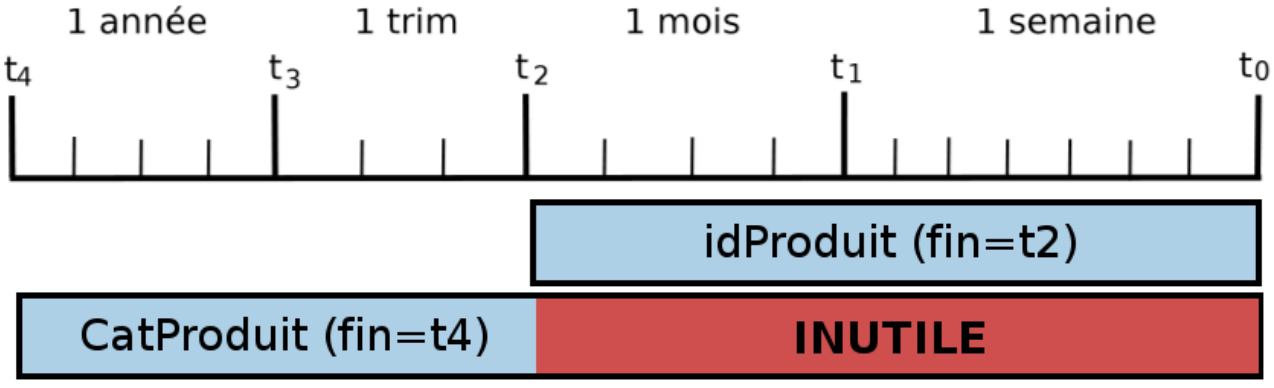
Les fonctions de précision, la genèse

1. Ne pas matérialiser ce qui n'est pas consulté
   $\Rightarrow$ Introduction d'une valeur *fin* pour chaque niveau



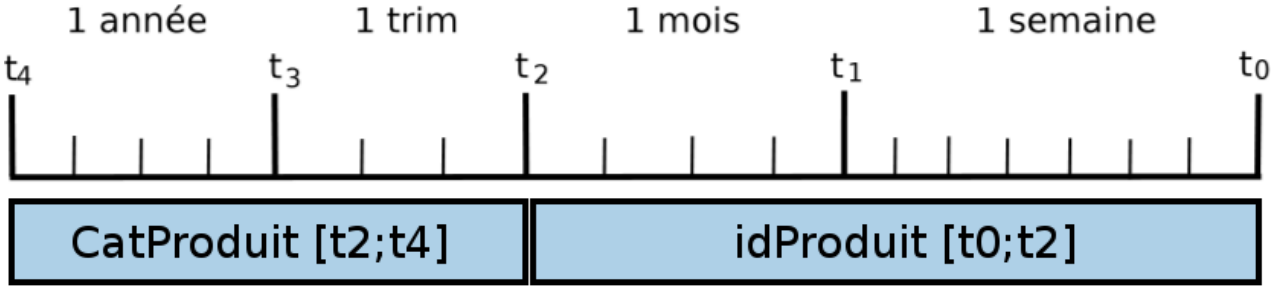
Les fonctions de précision, la genèse

1. Ne pas matérialiser ce qui n'est pas consulté
   ⇒ Introduction d'une valeur *fin* pour chaque niveau
2. Ne pas matérialiser ce qui est redondant
   ⇒ Introduction d'une valeur *debut* pour chaque niveau



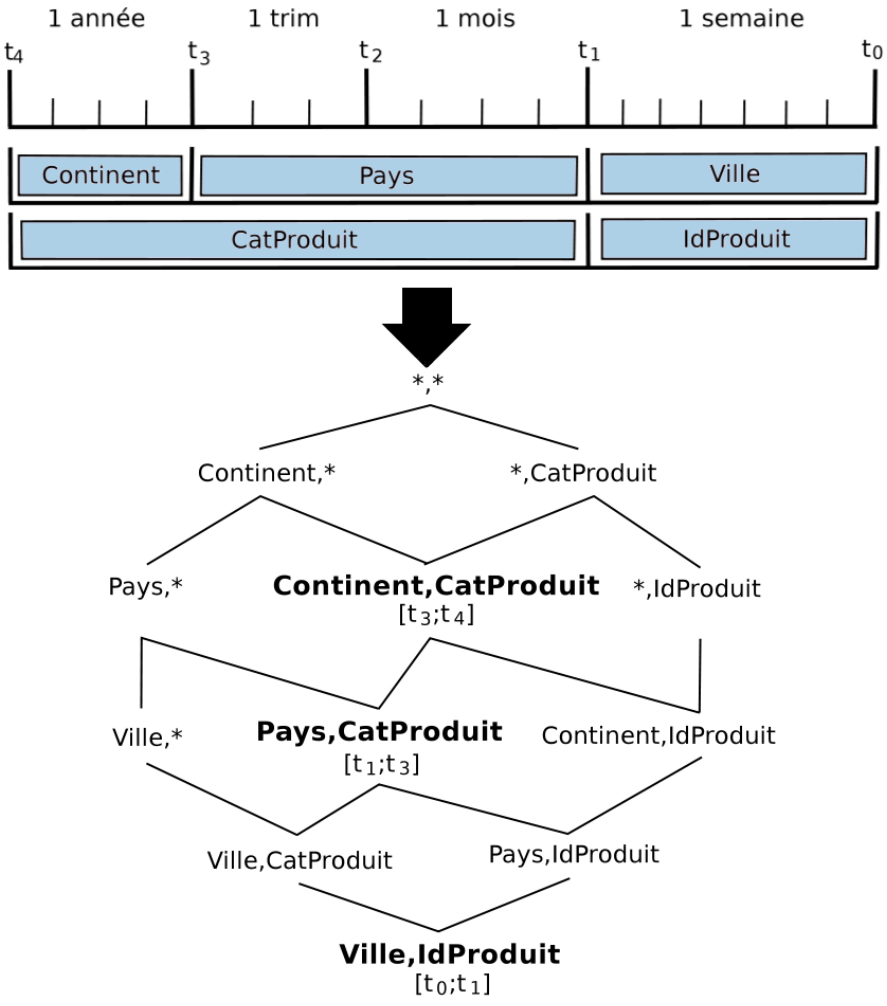
Les fonctions de précision, la genèse

1. Ne pas matérialiser ce qui n'est pas consulté
   $\Rightarrow$ Introduction d'une valeur *fin* pour chaque niveau
2. Ne pas matérialiser ce qui est redondant
   $\Rightarrow$ Introduction d'une valeur *debut* pour chaque niveau
3. Les fonctions de précision sont nées !
   ex : Précision(idProduit)=$[t_0; t_2]$



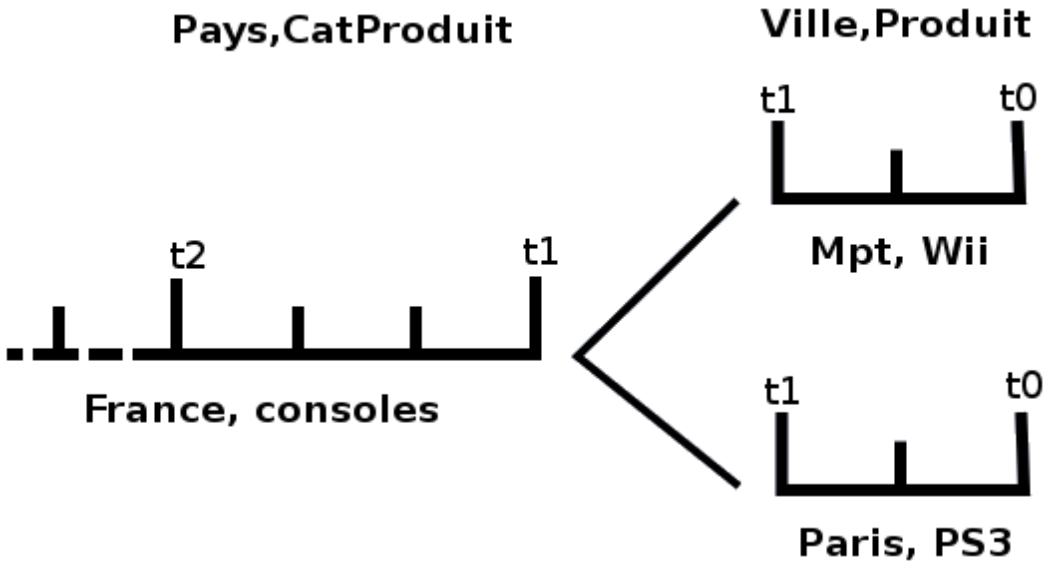
Combinaison des fonctions de précision

- Intersection : rapide et cohérent
- Peu de cuboïdes matérialisés
- Aucune redondance
- Ce qui n'est pas consulté n'est pas matérialisé



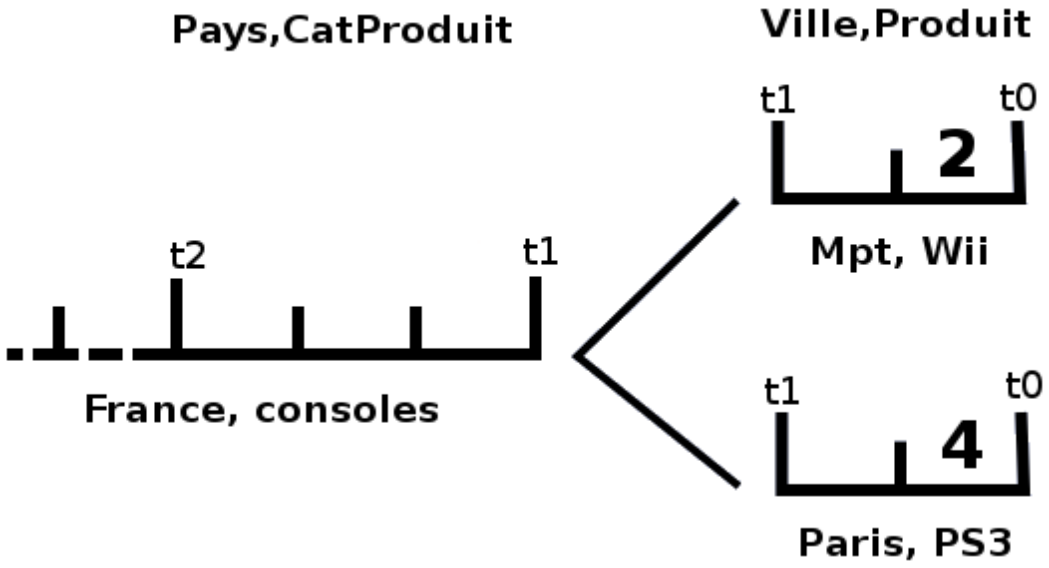
Insertion d'une valeur dans la structure

Tant que la fenêtre n'est pas pleine ⇒ processus "classique".



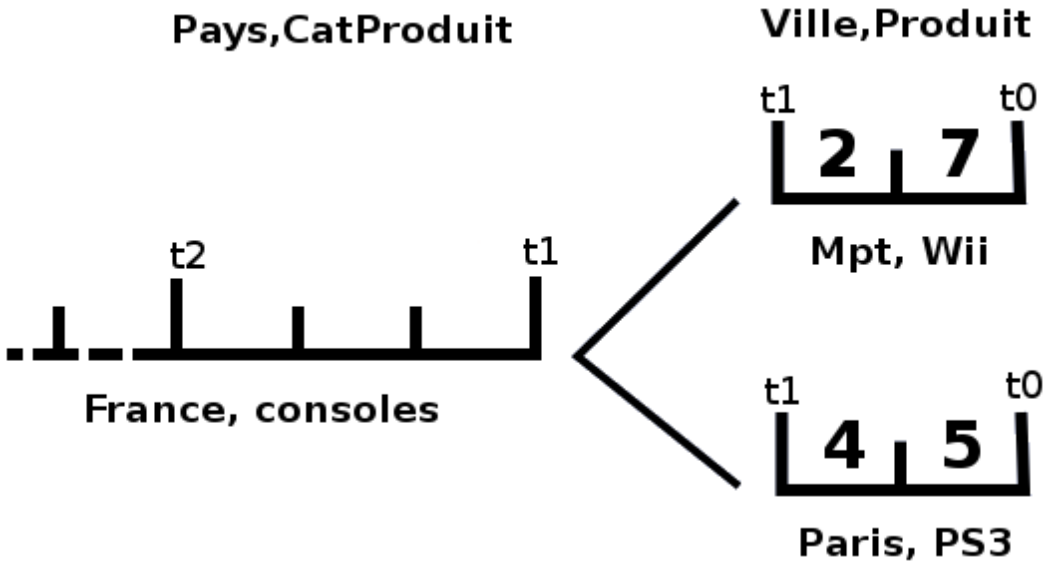
Insertion d'une valeur dans la structure

Tant que la fenêtre n'est pas pleine ⇒ processus "classique".



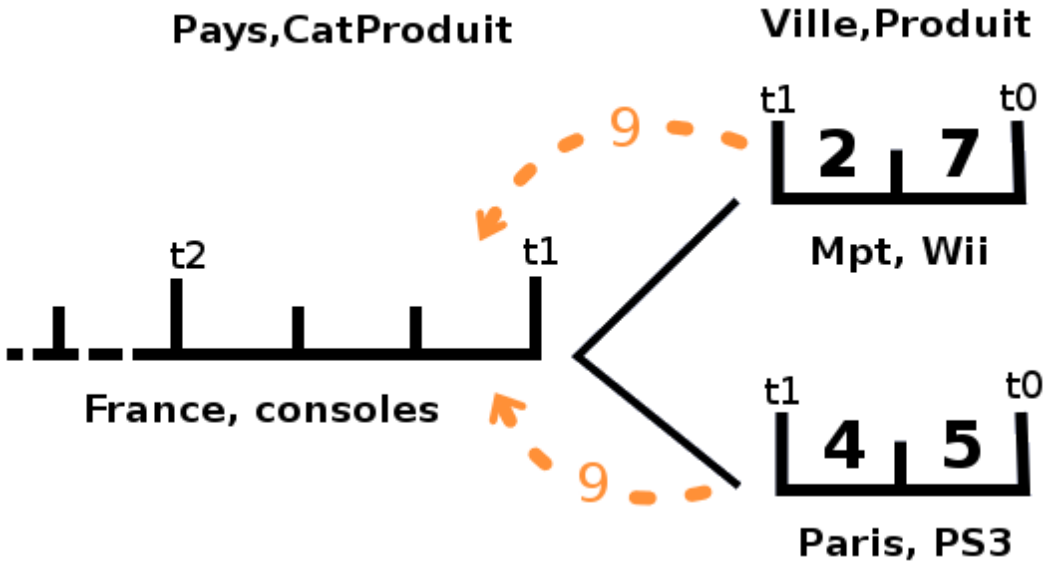
Insertion d'une valeur dans la structure

Tant que la fenêtre n'est pas pleine ⇒ processus "classique".



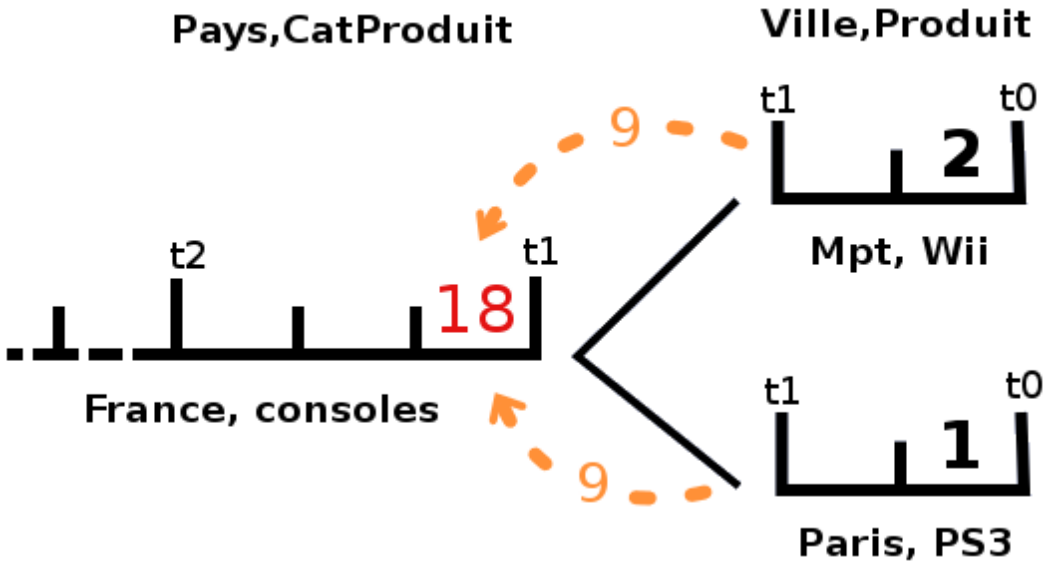
Insertion d'une valeur dans la structure

Fenêtre pleine + Passer au niveau supérieur ⇒ double agrégation.



Insertion d'une valeur dans la structure

Fenêtre pleine + Passer au niveau supérieur ⇒ double agrégation.



Importance de bien définir les fonctions de précision

Une bonne définition des fonctions de précision est cruciale pour la qualité de la structure
$\Rightarrow$ Nécessité d'inclure l'utilisateur dans le processus

Comment ?

1. **Manuellement** : dangereux dans un environnement multi-utilisateur
   Exemple : le directeur d'une chaîne nationale de magasins ne désire pas faire les mêmes analyses qu'un responsable de magasin.
2. **Automatiquement** : utilsation de logs de requêtes pour déterminer automatiquement quand un niveau d'abstraction n'est (presque) plus matérialisé

Automatiser la définition des fonctions de précision

Utilisation des logs

Le niveau 1 est interrogé 3 fois sur l'intervalle 2 ($[t_1; t_2]$).

Niveau	Intervalle	Frèq.
1	1	1
1	2	3
1	3	2
1	4	3
2	1	0
2	2	2
2	3	4
2	4	1
3	1	3
3	2	2
3	3	2
3	4	1

Automatiser la définition des fonctions de précision

Utilisation des logs

Le niveau 1 est interrogé 3 fois sur l'intervalle 2 ($[t_1; t_2]$).

Niveau	Intervalle	Frèq.
1	1	1
1	2	3
1	3	2
1	4	3
2	1	0
2	2	2
2	3	4
2	4	1
3	1	3
3	2	2
3	3	2
3	4	1

Automatiser la définition des fonctions de précision

Transformation des logs

Si l'on matérialise l'intervalle 2 sur le niveau 2 alors :

- On pourra répondre à 5 requêtes sur 7,
- On ne pourra pas répondre à 2 requêtes sur 7,
- Parmi les 5 requêtes satisfiables, 3 demanderont une généralisation

Niveau	% satisf.	% perte	général.
Intervalle 1			
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{1}$
1	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{0}{1}$
Intervalle 2			
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$
1	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{0}{3}$
Intervalle 3			
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$
1	$\frac{2}{8}$	$\frac{6}{8}$	$\frac{0}{2}$
Intervalle 4			
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$
1	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{0}{3}$

Automatiser la définition des fonctions de précision

Transformation des logs

Si l'on matérialise l'intervalle 2 sur le niveau 2 alors :

- On pourra répondre à 5 requêtes sur 7,
- On ne pourra pas répondre à 2 requêtes sur 7,
- Parmi les 5 requêtes satisfiables, 3 demanderont une généralisation

Niveau	% satisf.	% perte	général.
Intervalle 1			
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{1}$
1	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{0}{1}$
Intervalle 2			
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$
1	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{0}{3}$
Intervalle 3			
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$
1	$\frac{2}{8}$	$\frac{6}{8}$	$\frac{0}{2}$
Intervalle 4			
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$
1	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{0}{3}$

Automatiser la définition des fonctions de précision

Suppression des alternatives trop imprécises

On fixe un seuil σ, et on supprime les lignes du tableau où %perte$>\sigma$

Ici, σ=30%

Niveau	% satisf.	% perte	général.
Intervalle 1			
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{1}$
1	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{0}{1}$
Intervalle 2			
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$
1	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{0}{3}$
Intervalle 3			
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$
1	$\frac{2}{8}$	$\frac{6}{8}$	$\frac{0}{2}$
Intervalle 4			
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$
1	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{0}{3}$

Automatiser la définition des fonctions de précision

Suppression des alternatives trop imprécises

On fixe un seuil σ, et on supprime les lignes du tableau où %perte$>\sigma$

Ici, σ=30%

Niveau	% satisf.	% perte	général.
Intervalle 1			
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{1}$
1	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{0}{1}$
Intervalle 2			
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$
1	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{0}{3}$
Intervalle 3			
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$
1	$\frac{2}{8}$	$\frac{6}{8}$	$\frac{0}{2}$
Intervalle 4			
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$
1	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{0}{3}$

Automatiser la définition des fonctions de précision

Suppression des alternatives trop imprécises

On fixe un seuil σ, et on supprime les lignes du tableau où %perte$>\sigma$

Ici, σ=30%

Niveau	% satisf.	% perte	général.
Intervalle 1			
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$
Intervalle 2			
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$
Intervalle 3			
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$
Intervalle 4			
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$

Automatiser la définition des fonctions de précision

Fonction de coût

Objectif : trouver le meilleur compromis entre la précision et la généralisation

Solution : somme entre le pourcentage de perte et la généralisation

αxPerte$+(1-\alpha)$xGeneralisation

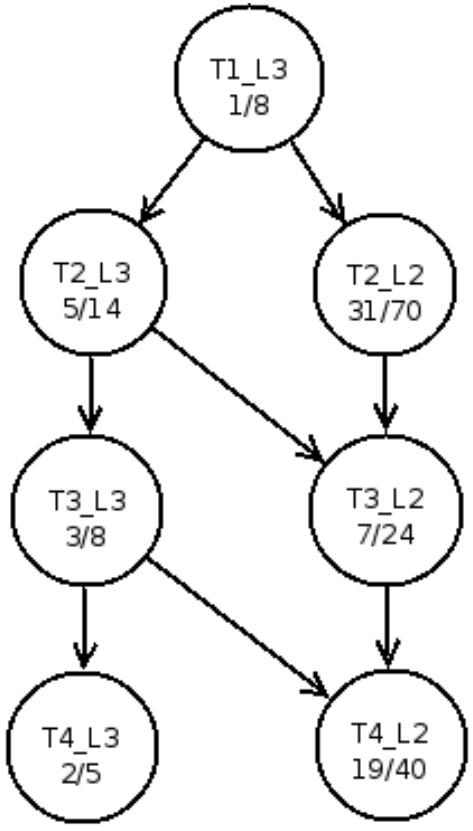
Niveau	% satisf.	% perte	général.	coût
Intervalle 1				
3	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{1}{4}$	$\frac{1}{8}$
Intervalle 2				
3	$\frac{7}{7}$	$\frac{0}{7}$	$\frac{5}{7}$	$\frac{5}{14}$
2	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{3}{5}$	$\frac{31}{70}$
Intervalle 3				
3	$\frac{8}{8}$	$\frac{0}{8}$	$\frac{6}{8}$	$\frac{3}{8}$
2	$\frac{6}{8}$	$\frac{2}{8}$	$\frac{2}{6}$	$\frac{7}{24}$
Intervalle 4				
3	$\frac{5}{5}$	$\frac{0}{5}$	$\frac{4}{5}$	$\frac{2}{5}$
2	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{4}$	$\frac{19}{40}$

Automatiser la définition des fonctions de précision

Transposition sous forme de graphe et application d'un algorithme de plus court chemin

Un arc (n,n') existe ssi :

- $T_{n'} = T_n + 1$
- $L_{n'} \geq L_n$

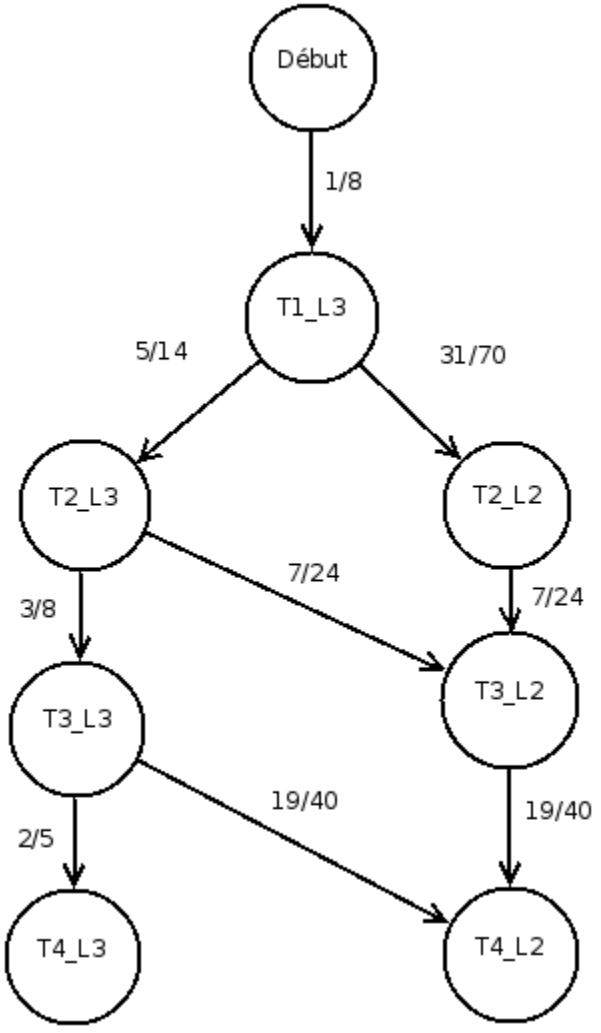


Automatiser la définition des fonctions de précision

Transposition sous forme de graphe et application d'un algorithme de plus court chemin

Un arc (n,n') existe ssi :

- $T_{n'} = T_n + 1$
- $L_{n'} \geq L_n$



Automatiser la définition des fonctions de précision

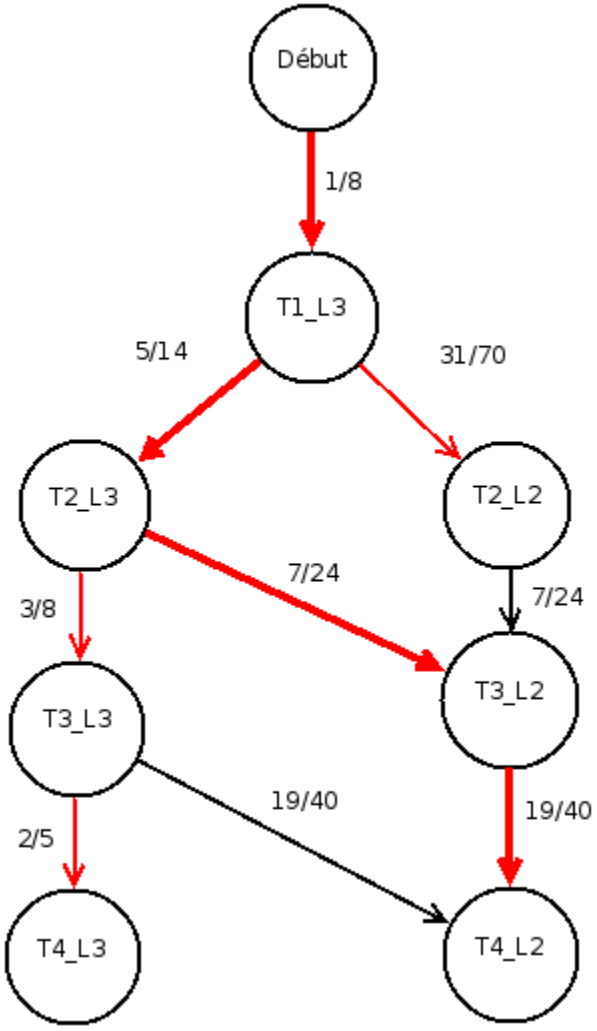
Transposition sous forme de graphe et application d'un algorithme de plus court chemin

Un arc (n,n') existe ssi :

- $T_{n'} = T_n + 1$
- $L_{n'} \geq L_n$

Résultats :

- Precision(L3)=$[t_0; t_2]$
- Precision(L2)=$[t_2; t_4]$



Interrogation

Problématiques ?

- Comment déterminer si une réponse précise à une requête est possible ?
- Le cas échéant, comment réagir ?

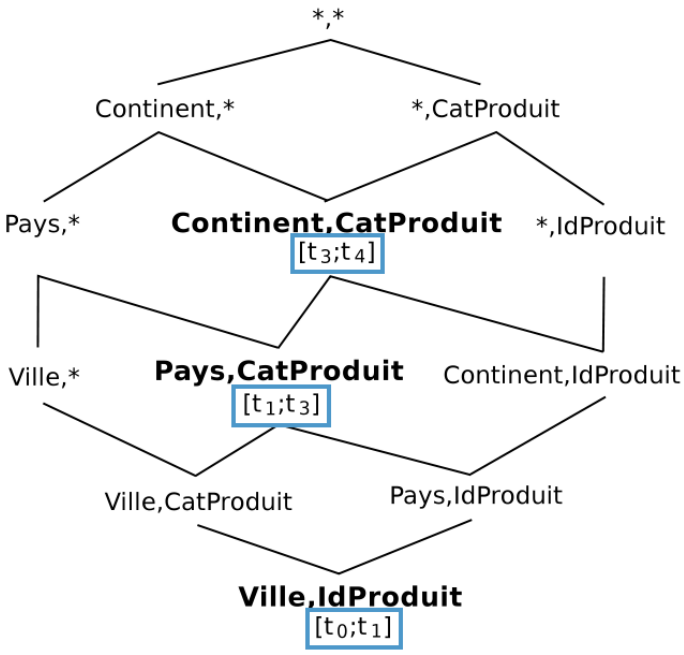
Méthodes

1. Déterminer la satisfaisabilité d'une requête grâce aux métadonnées (fonctions de précision)
2. Proposer des requêtes *alternatives* [**?**] si une réponse précise est impossible

Illustration

Lister les ventes par pays et par produits entre t_0 et t_4.

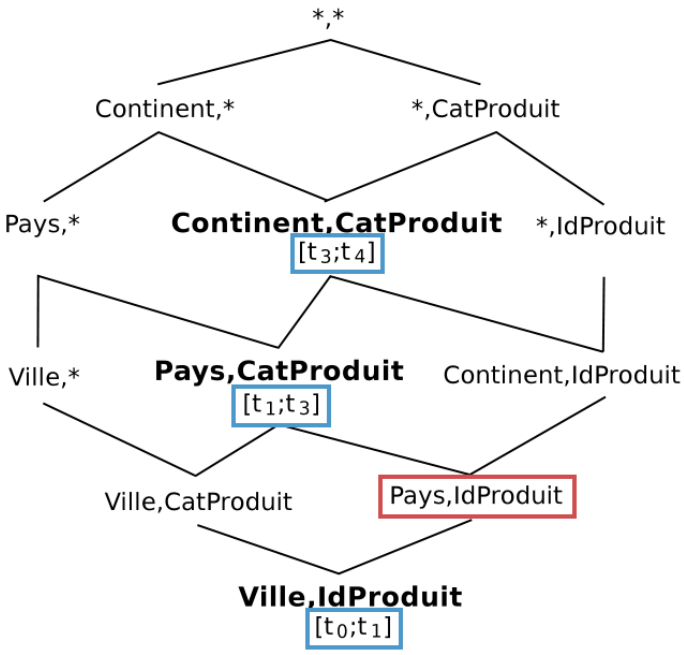
Exhiber l'intervalle de temps concerné par la requête



Illustration

Lister les ventes par pays et par produits entre t_0 et t_4.

Identifier le cuboïde concerné par la requête et déterminer la satisfaisabilité.



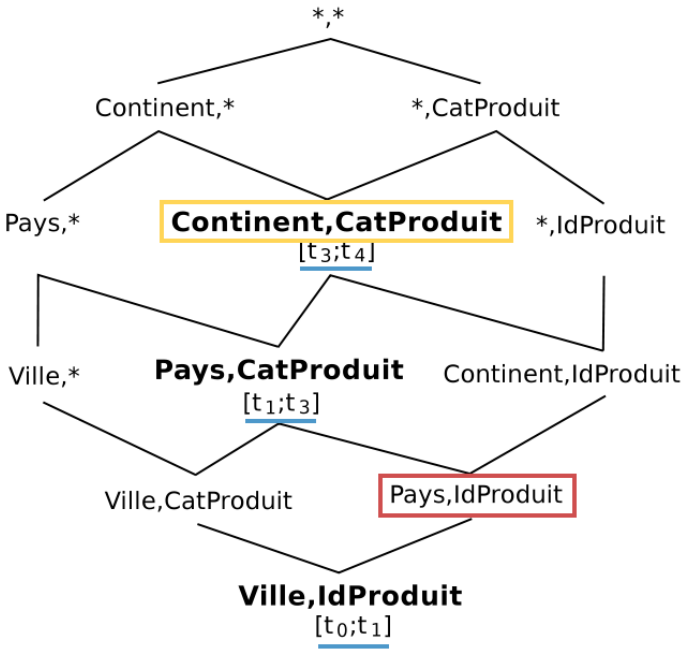
La requête n'est pas satisfiable.

Illustration

Lister les ventes par pays et par produits entre t_0 et t_4.

Proposer une requête alternative

Lister les ventes par continents et par catégories de produits entre t_0 et t_4.

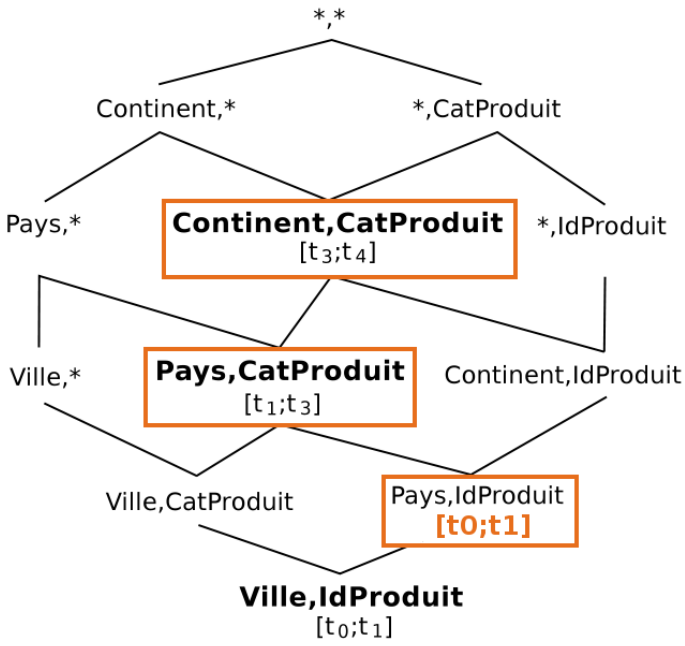


Illustration

Lister les ventes par pays et par produits entre t_0 et t_4.

Si la requête alternative est refusée, proposer une réponse verbeuse

- La liste des ventes par continents et catégories de produits entre t_3 et t_4
- La liste des ventes par pays et catégories de produits entre t_1 et t_3
- La liste des ventes par pays et produits entre t_0 et t_1



Plan

Expérimentations

Objectifs

- Une structure compacte
- Eviter explosion mémoire vive
- Temps d'insertion d'un batch borné

Protocole expérimental

1. Comparaison de la matérialisation avec StreamCube
2. Evaluation de la consommation de la mémoire vive et du temps d'insertion d'un batch sur des jeux de données synthétiques difficiles
3. Application de la proposition sur un jeu de données réelles (issues de sondes réseaux)

Comparaison avec StreamCube

Jeux de données synthétiques (générateur de batchs)

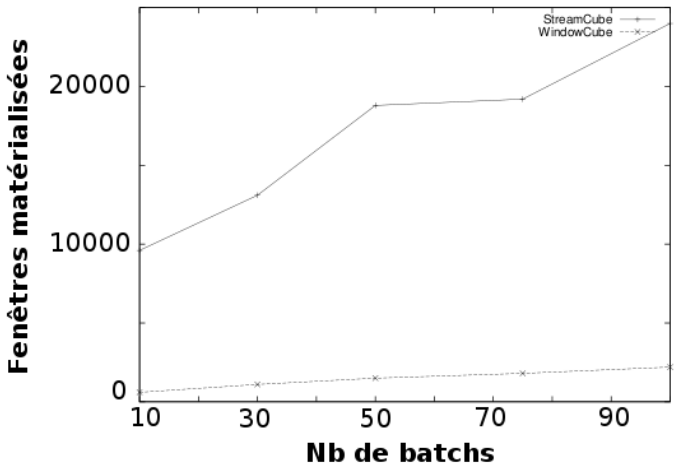


FIG.: Fenêtres matérialisées en fonction du nb. de batchs (D5C5L3T1)

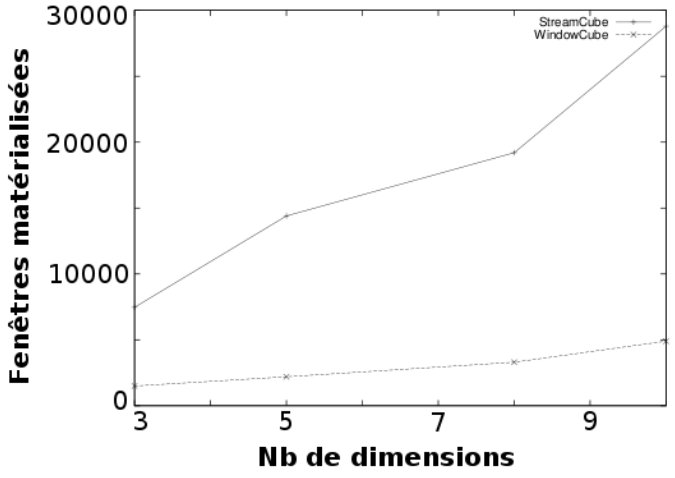


FIG.: Fenêtres matérialisées en fonction du nb. de dimensions (C5L3B50T1)

Entre 6 et 10 fois moins de fenêtres matérialisées tout en gardant une qualité de réponse satisfaisante.

Application sur des jeux de données synthétiques

Paramètres testés : fonctions de précision, dimensions, degré des dimensions, profondeur des dimensions.

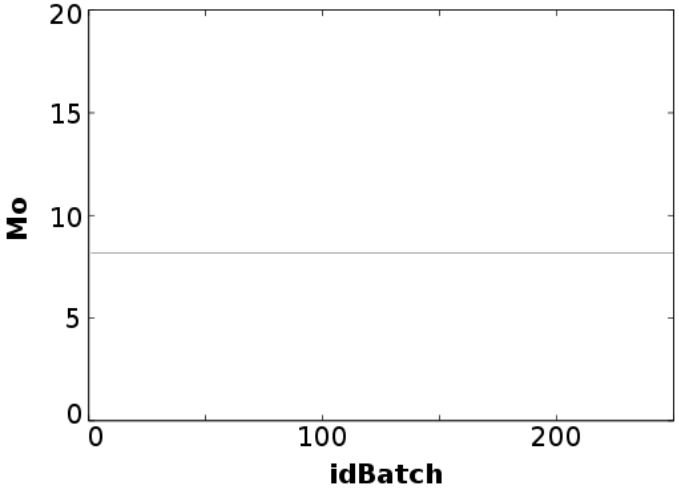


FIG.: RAM consommée par batchs (D5L3C5B250T20)

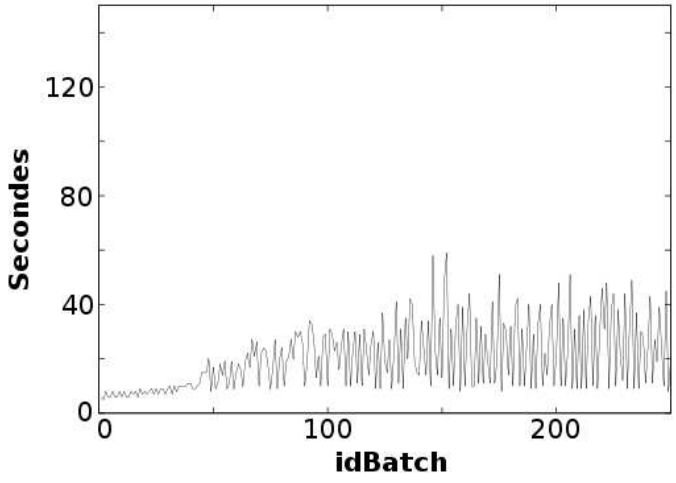


FIG.: Temps d'insertion par batchs (D5L3C5B250T20)

- Mémoire vive faible et stable
- Un temps d'insertion borné à 60 secondes

Application sur un jeu de données réelles

Données issues de sondes réseaux, 13 dimensions, 204 batchs de 20K n-uplets.

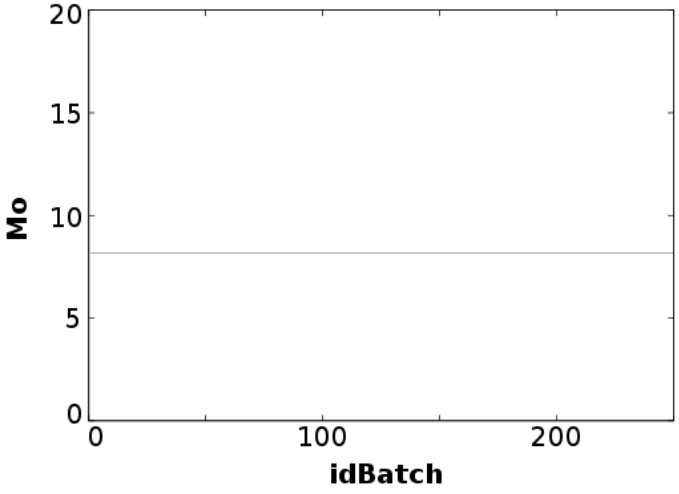


FIG.: RAM consommée par batchs

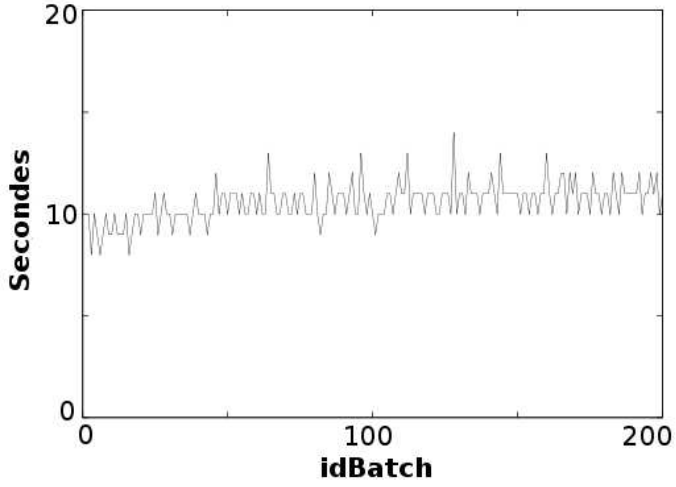


FIG.: Temps d'insertion par batchs

- Mémoire vive faible et stable
- Un temps d'insertion borné à 15 secondes

Plan

Conclusion

Une méthode de résumé de flots de données multidimensionnelles

1. Minimiser la taille de la structure ✓
   - Combinaison des fonctions de précision
   - Expérimentations satisfaisantes
2. Maximiser la qualité des réponses ✓
   - Définition automatique des fonctions de précision pour borner l'imprécision
   - Méthode de gestion de l'imprécision

Perspectives

- Mise en place de techniques de fouille de données
- Proposer un langage d'interrogation continu
- Réfléchir à l'aspect multi-flots (panne, étude de corrélations)
- Visualisation

Merci de votre attention

J. Han, Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang, and Y. D. Cai.

Stream cube : An architecture for multi-dimensional analysis of data streams.

Distrib. Parallel Databases, 18(2), 2005.

R. B. Messaoud.

Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes.

PhD thesis, Ecole Doctorale Sciences Cognitives de l'Université Lumière Lyon 2, 2006.

T. B. Pedersen, C. S. Jensen, and C. E. Dyreson.

Supporting imprecision in multidimensional databases using granularities.

In *SSDBM '99 : Proceedings of the 11th International Conference on Scientific on Scientific and Statistical Database Management*, page 90, Washington, DC, USA, 1999. IEEE Computer Society.